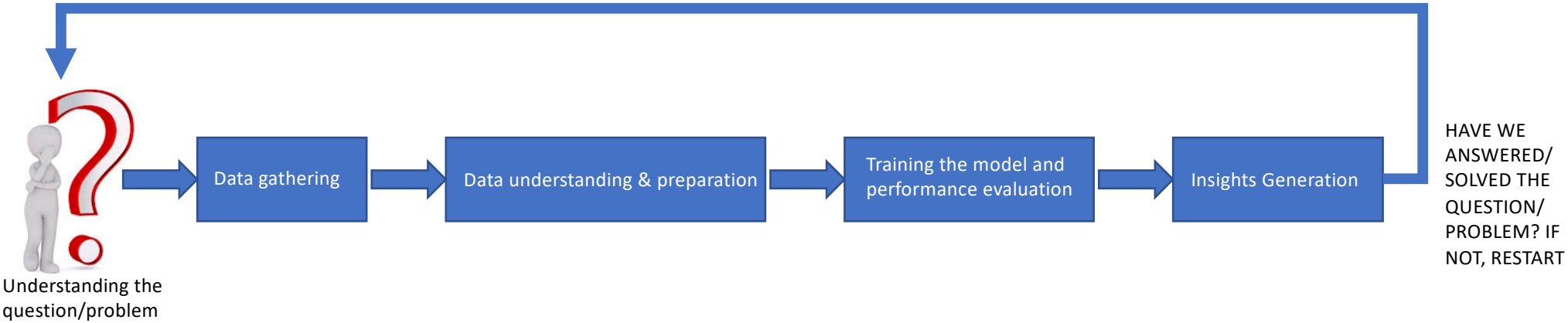
PROJECT 1: Application of clustering to a real use case: **Customer segmentation**

THE GOAL:

Analyze and understand the main characteristic of **a customer base for marketing purposes** by means of clustering techniques

METHOD:

During this project we will follow the common end-to-end Machine Learning process: from understanding the problem, data gathering and cleaning, exploratory data analysis, feature engineering and finally, training and prediction.



PROJECT 1: Application of clustering to a real use case: **Customer segmentation**

Data gathering

Understanding the question/problem

UNDERSTANDING THE PROBLEM:

Recently, we have joined a data scientist and AI engineering team of a telco company. This team is supporting decision making process of several internal areas as marketing and customer care.
Our marketing colleagues are planning to launch a new commercial campaign for a new mobile tariff. As input for the tariff design, the Marketing product managers need to know the **pattern** or **stereotypes** of our current customers: i.e. the main customers' sectors (Industry, Services, Agriculture, ...), number of employees, Turnover, which products they consume, how much they spend in telco services, etc.

DATA GATHERING: The data extracted from our Data WareHouse is:

- *Company_id*: It's an integer that identifies any company. Each value is unique for every company.
- *Reference_date_month*: It refers to the month corresponding to the customer information
- *Product*: Name level 1 of the product purchased by the customer
- *Sub_product*: Name level 2 of the product purchased by the customer
- *Type_ID_Coverage_GSM*: type of coverage for GSM (2G) network: indoor, outdoor or no-coverage
- *Type_ID_Coverage_UMTS*: type of coverage for UMTS (3G) network: indoor, outdoor or no-coverage
- *Type_ID_Coverage_LTE*: type of coverage for LTE (4G) network: indoor, outdoor or no-coverage
- CNT_EMPLOYEE: Number of employees of the company

- *Sector*: It's an integer that identifies the sector of the company's activity
- *Sub_sector*: It's an integer that identifies the sub_sector of the company's activity
- *Turnover*: The annual incomes of the company
- *ZIP*: Postal code where the company is placed
- *Data_usage*: Number of Gigabits for this product and sub_product and Reference_date_month
- *Voice_usage*: Number of minutes for this product and sub_product and Reference_date_month
- *Monthly_expense*: Euros expended in this product and sub_product and Reference_date_month
- *N_lines*: Number of mobile lines correspoding to this product and sub_product and Reference_date_month

PROJECT 1: Application of clustering to a real use case: **Customer segmentation**

Data understanding & preparation

DATA UNDERSTANDING AND PREPARATION

Once we know the problem to solve, the next stage is to have a clear understanding of the data we have extracted and to prepare it before clustering. In particular, we will:

- List and verify the type of each variable (object, float, int…). Identify variables with nulls. Measure the memory usage
- Eliminate rows with nulls in order to have a dataset 100% fulfilled
- Aggregate rows with monthly expense per customers in order to have just 1 sample per customers
- Exploratory Data Analysis to understand main statistics (mean, standard deviation, min&max values and 25%-50%-75% quartiles) and distribution of the most relevant variables or features as data usage, voice usage, monthly expense and number of lines
- Plot several graphs in order to identify how variables are related between them. In particular:
    - correlation matrix
    - 2D and 3D scatter plots between data usage, voice usage and monthly expense

Once this part of the Project is done, we should achieve a deep knowledge about the data. Besides, the dataset will have been processed to be ready to apply the clustering algorithms to solve the business problem.

PROJECT 1: Application of clustering to a real use case: **Customer segmentation**

Training the model and performance evaluation

K-MEANS: TRAINING THE MODEL AND PERFORMANCE EVALUATION

Firstly, we will code our own Kmeans algorithm[1]. As the dataset has a high number of features, we will select voice usage, data usage and monthly expense variables to fit the clusters.

Once the clustering is done, we need to understand the output. Visualization of 2D and 3D scatter plots is a excellent technique to evaluate the clustering output.

To check if our Kmeans algorithm works properly, we will use the Sklearn's Kmeans function[2] to cluster the dataset. We will compare the 2D and 3D plots from the Sklearn clustering and ours.

Finally, as part of any Machine Learning Project, we need to calculate the perfomance of our model. For Kmeans, we will 1) estimate the optimal K value through the Elbow method and 2) calculate the sihouette score for several values of K

(1) You can use the one you developed in P1 if it works with datasets with more than 2 dimensions.

(2) The Data Science and Pythonist Community is huge and have developed Sklearn, a great Python library that includes the most relevant Machine Learning algorithms. Sklearn facilitates the training process of an algorithm and the posterior application in the prediction stage.

PROJECT 1: Application of clustering to a real use case: **Customer segmentation**

Insights Generation

K-MEANS: INSIGHTS GENERATION

It is time to answer the initial request from the marketing department: i.e. describe the different groups of customers that purchased our products.

To do it, we:

- Apply the Sklearn' Kmeans function for a new dataset formed by number of employees, turnover, voice and data usage and montly expense variables. We will use K=3.
- For every cluster, we will describe the main statistics for each variables
- Finally, we will compare the statistics of 2) of each cluster to understand how are the companies in each group. Some questions to answer:
  - Which is the cluster with the highest voice usage?
  - Which is the cluster with the highest data usage?
  - Customers in the cluster with bigger companies (i.e. bigger number or employes and turnover) use to spend more than the others customers?
  - As a part of the data scientist team, which is your recommended cluster of customers to sell a new mobile tariff with unlimited data traffic? And for a new mobile tariff with unlimited voice traffic?

PROJECT 1: Application of clustering to a real use case: **Customer segmentation**

MIXTURE OF GAUSSIANS: TRAINING AND INSIGHTS GENERATION

As we have learned in this course, Mixture of Gaussians is a good solution when there is covariance between variables.
We are going to train a MoG model using the Sklearn's library and repeat the process. In particular:

- Execute the Mixture of Gaussians function (with number of components=3) to the dataset with Voice_usage, Data_usage and Monthly_expense variables.
- Visualize the 2D and 3D scatter plots
- Calculate the performance, i.e. the sihouette score, for different number of K
- Compare the main statistics of each cluster considering employees, turnover, voice and data usage and montly expense variables. How are the companies that belong to each cluster?

PROJECT 1: Application of clustering to Customer segmentation

THINKING "OUT-OF-THE BOX"

Clustering is a very useful technique to group or segment a customer base. However, the possibilities of real use cases and applications where unsupervised methodologies can play a key role are huges. In the following exercise, you will use these techniques to plan the location of a fast food restaurant, a school and a laundry according to description of the neighborhoods of Barcelona.