

# Content

- **Case Study Intro**
- **DBSCAN Intro**
  - Density and Dense Region
  - Min Points and Eps
  - Core point
  - Border point
  - Noise point
  - Density Edge
  - Density Connected Points
- **DBSCAN Algo**
- **Hyperparam tuning**
- **Adv and Disadv of DBSCAN**
- **Implementing DBSCAN**
- **Intro to Anomaly detection**
  - Distribution based
  - Elliptic Envelope

## ▼ Case Study: Customer Grouping with DBSCAN

In today's lecture, we'll try to segment customers based on 'Wholesale Customers Data'.

### **Q. What is meant by customer segmentation?**

- Customer segmentation is the process by which you divide your customers up based on common characteristics – such as demographics or behaviours, so you can market to those customers more effectively.
- Almost all the companies in today's world, one way or other, make use of customer segmentation for marketing.

### **Q. But, how would we segment the customers based on some numbers present in the data?**

- So far, we have seen K-Means which is a centroid based algorithm.
- We saw Agglomerative Clustering which is hierarchical system.

- Today we will see an overview (very high level) for 2 more clustering algorithms. The goal for this lecture is not to go into detail about these algorithms but to just get intuitive understanding of how they work, since they are not as popular in the community.
- We'll use these algorithms to do the segmentation of the customers on 'Wholesale Customers Data'.
- First, we will start with DBSCAN which is more of density based; i.e. the clustering should be done on the basis of how dense a neighborhood around a point.

Before jumping in, let's first understand and import the data, and try to visualize it on a plot.

## ▼ Dataset - Wholesale Customers Data

The dataset used for this case study contains the following parameters:

1. **Fresh**: annual spending (some monetary unit) on fresh products (Continuous);
2. **Milk**: annual spending (some monetary unit) on milk products (Continuous);
3. **Grocery**: annual spending (some monetary unit) on grocery products (Continuous);
4. **Frozen**: annual spending (some m.u.) on frozen products (Continuous)
5. **Detergentss\_Paper**: annual spending (some m.u.) on detergents and paper products (Continuous)
6. **Delicassen**: annual spending (some m.u.) on and delicatessen products (Continuous);
7. **Channel**: customers Channel - (Two possible values: Horeca (Hotel/Restaurant/Cafe) or Retail channel) (Nominal)
8. **Region**: customers Region (Three possible values: Lisnon, Oporto or Other) (Nominal)

```
1 !gdown 10ZdMFvfhGeXAa8xR0J00kHIH9y9m9uL
```

```
  Downloading...
  From: https://drive.google.com/uc?id=10ZdMFvfhGeXAa8xR0J00kHIH9y9m9uL
  To: /content/wholesaledata.csv
  100% 15.0k/15.0k [00:00<00:00, 19.9MB/s]
```

```
1 from sklearn.cluster import DBSCAN
2 from sklearn.preprocessing import StandardScaler
3 import numpy as np
4 import pandas as pd
5 import matplotlib.pyplot as plt
6 plt.rcParams["figure.figsize"] = (16,12)
```

```
1 df = pd.read_csv('./wholesaledata.csv')
2 df.info()
```