# Outliers

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

# Handling outliers using DBSCAN method for multidimensional data

DBSCAN is a density-based clustering algorithm, it is focused on finding neighbors by density (MinPts) on an ‘n-dimensional sphere’ with radius ɛ. A cluster can be defined as the maximal set of ‘density connected points’ in the feature space.

DBSCAN then defines different classes of points:

* **C*ore point****:* **A** is a core point if its neighborhood (defined by ɛ) contains at least the same number or more points than the parameter MinPts.
* ***Border point***: **C** is a border point that lies in a cluster and its neighborhood does not contain more points than MinPts, but it is still ‘*density reachable’*by other points in the cluster.
* ***Outlier***: **N** is an outlier point that lies in no cluster, and it is not ‘*density reachable’* nor ‘*density connected’* to any other point. Thus, this point will have “his own cluster”.

A cluster satisfies two properties:

1. All points within the cluster are mutually density-connected.
2. If a point is density-reachable from any point of the cluster, it is part of the cluster as well.

Again, the first step is scaling the data since the radius **ɛ** will define the neighborhoods along with **MinPts**. After scaling the feature space, is time to choose the spatial metric on which DBSCAN will perform the clustering. The metric must be chosen depending on the problem, a Euclidean metric works well for 2 or 3 dimensions, the Manhattan metric can also be useful when dealing with higher dimensional feature spaces 4 or more dimensions. Then, the parameter eps (ɛ) must be chosen accordingly to perform clustering. If ɛ is too big many points will be density connected, if it is too small the clustering will result in many meaningless clusters.

Outliers (noise) will be assigned to the -1 cluster. After tagging those instances, they can be removed or analyzed.

# Are all outliers useless data?

We have to consider a couple of factors before deciding whether outliers are useless and should be dropped or not:

1. If it is obvious that the outlier is due to incorrectly entered or measured data, we should drop the outlier.
2. If the outlier does not change the results but does affect assumptions, we may drop the outlier.
3. Usually, the outlier affects both results and assumptions.  In this situation, it is notlogical to simply drop the outlier. Model should be analyzed both with and without outliers to decide.
4. If the outlier creates a strong association, we should drop the outlier and should not report any association from our analysis.