Unsupervised Learning

Learning is said to be unsupervised if the experience/dataset doesn't come with the responses.

The agent's experience doesn't come with responses concerning the correctness of performing the task.

Unsupervised learning agents learn the structure, similarities and differences of the input

Unsupervised Learning Problems

Clustering

Most of unsupervised learning problems are clustering problems. These are problems in which the agent should be able to distinguish different clusters from the dataset based only on the features.

Unsupervised Learning Problems

Density Estimation

Density estimation is a technique usually used by statisticians which is used to estimate the underlying probability density functions of a given dataset. Examples of these techniques are histograms, kernel density estimations and etc [@Bishop2006].

Unsupervised Learning Problems

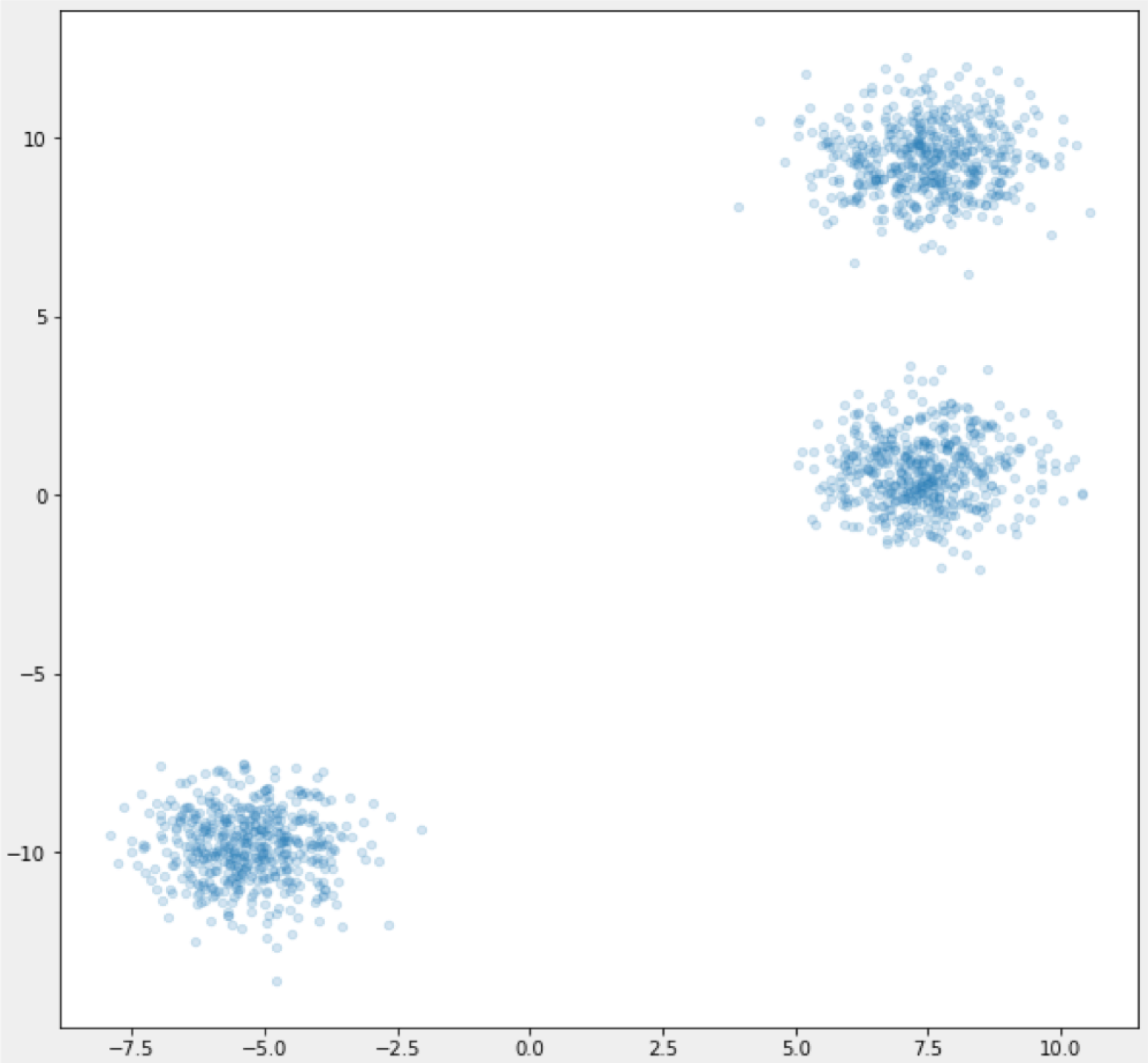
Projection Pursuit

This is another technique which is finds the most *interesting* projection of a given multidemnsional data. A projection is some form of transformation from one vector space to another. One of it's common uses is to reduce some high dimensional data into low dimensional data. This is done by projecting a high dimensional vector space to low dimensional vector space [@Friedman&Tukey1974].

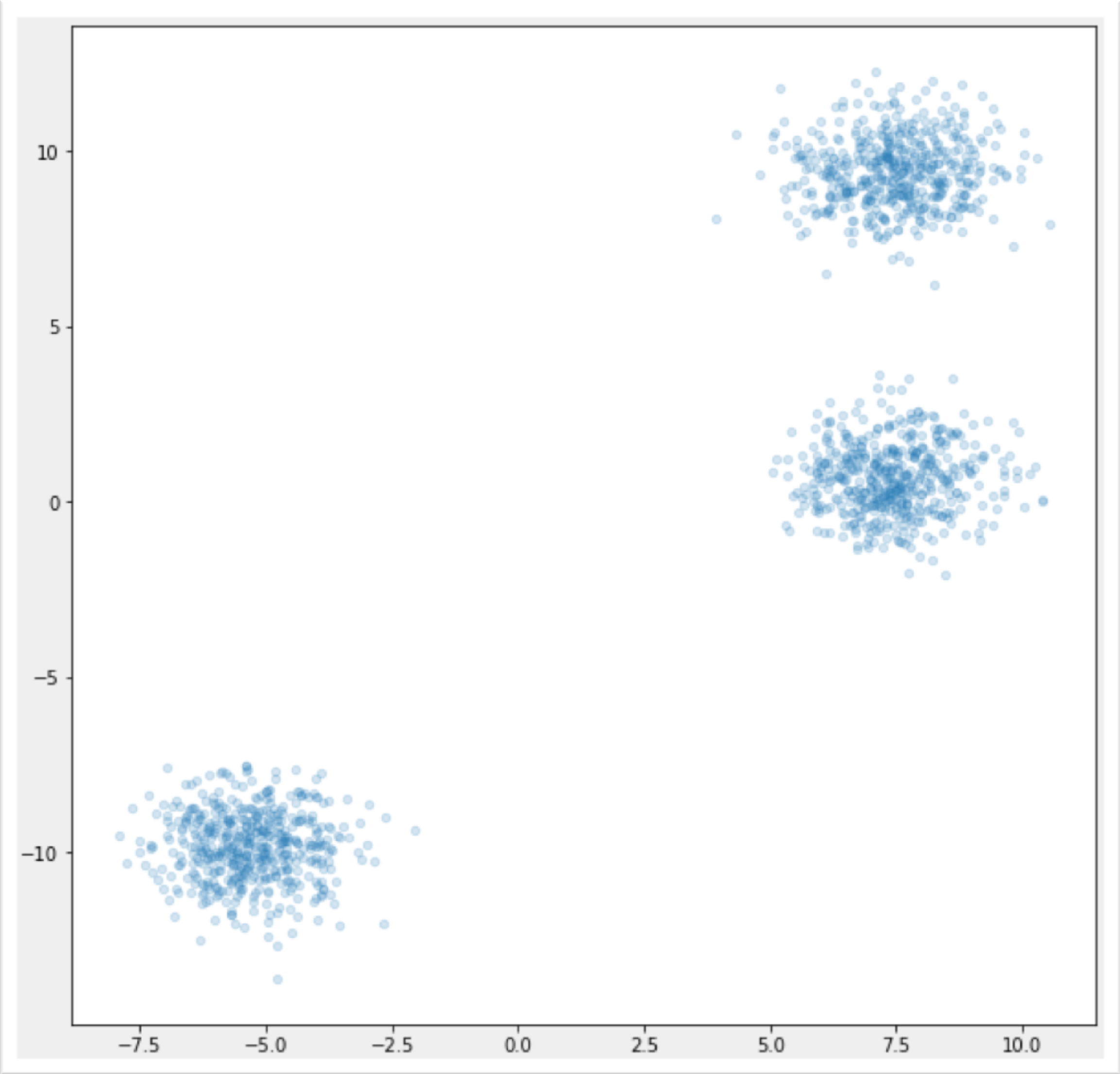
Clustering

k-Means Clustering

The simplest form of clustering in a dataset is a dataset that contains some k gaussian blobs, each with equal variances across all features.


10
5
0
−5
−10
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

It is easy to cluster these types of datasets, the agent only needs to find 3 centroids (center of a blob) which would define how the data is clustered. This can easily be done using the k-Means model.



K-means clustering divides a dataset X into k disjoint clusters c. Each cluster c is defined by a centroid μ_c. A sample $x^{(i)}$ is said to be of cluster c if the examples is closest to centroid μ_c. Or:

$$C^{(i)} = \text{argmin}_{\mu_c \in \{\mu_1, \mu_2, \ldots, \mu_k\}} ||x^{(i)} - \mu_c||$$

This means that the objective of K-means clustering is finding the optimal set of centroids, $\{\mu_1, \mu_2, \ldots, \mu_k\}$, or the set of centroids that will yield the minimum within-cluster sum of squared distances. This measure is also known as the **inertia** denoted by I [@scikitlearn2017].

$$I = \sum_{i=1}^{m} ||x^{(i)} - C^{(i)}||^2$$

Lloyd's Algorithm

After initializing a set k of centroids $\{\mu_1, \mu_2, \dots, \mu_k\}$ (usually initialized randomly):

- **Assignment Step**. Assign each example to the cluster nearest to it

  For each $i \in \{1,2,\dots,m\}$

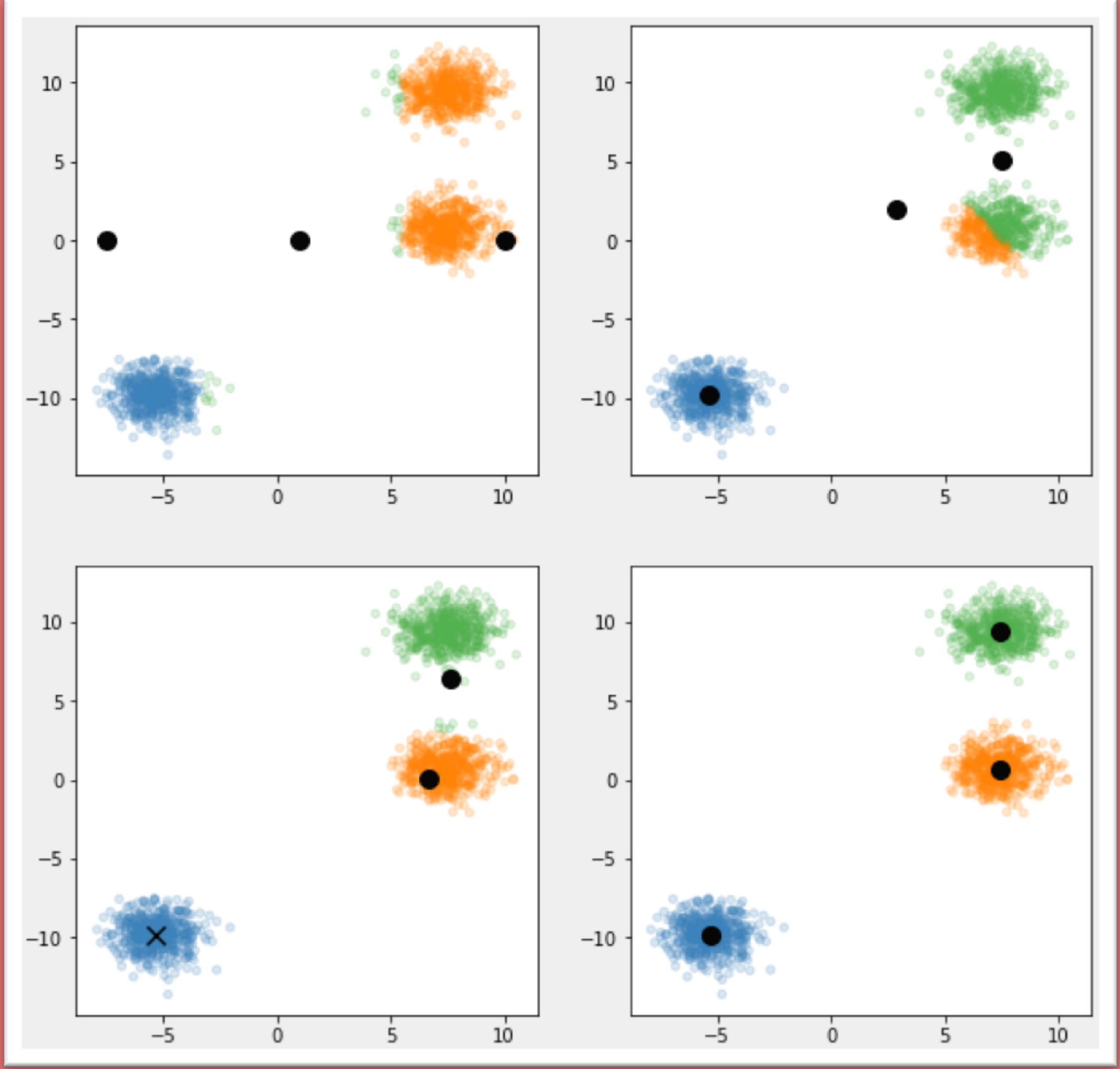
  $$C^{(i)} := \min_{c\in\{1,2,\dots k\}} ||x^{(i)} - \mu_c||$$

- **Update Step**. Calculate new centroids as the within-cluster mean of all examples

  For each $k \in \{1,2,\dots,k\}$

  $\mu_c := \frac{1}{s(c)} \sum_{i=0}^{s(c)} x^{(i)}$ (where $s(c)$ is the number of examples assigned to centroid c)

After a sufficient amount of iterations, switching between the assignment and update step, the algorithm will converge into the optimal set of centroids.

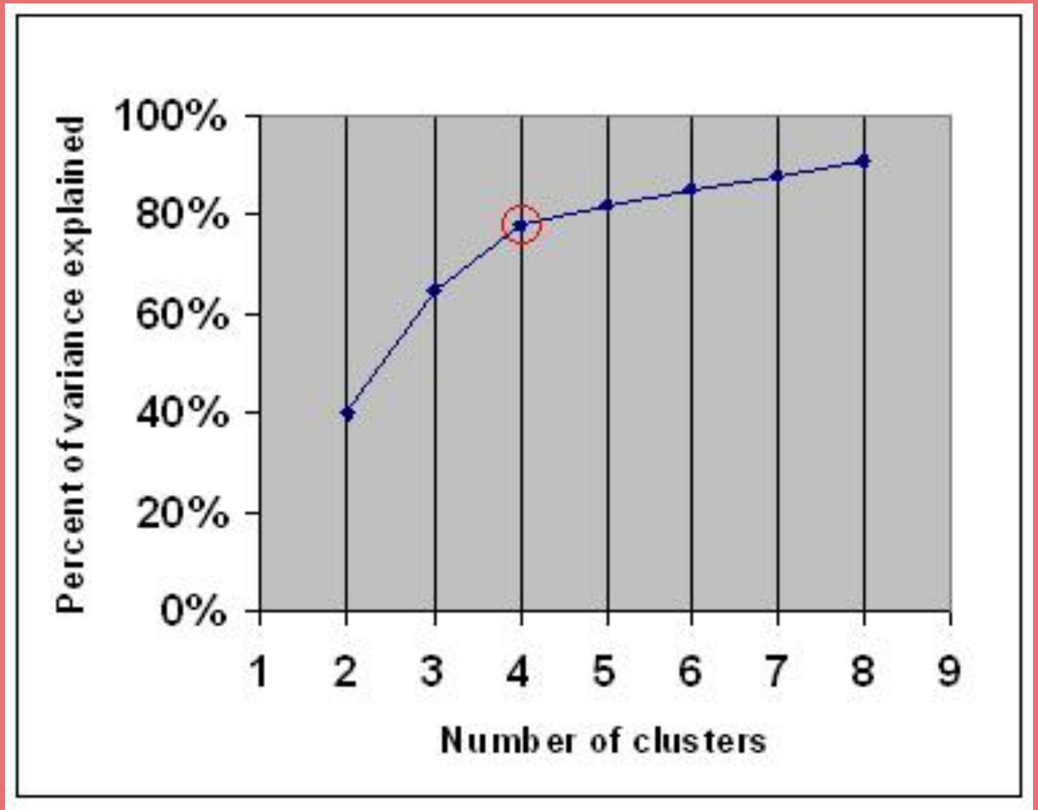


Limitations of k-Means Clustering Models

Choosing the value for k.

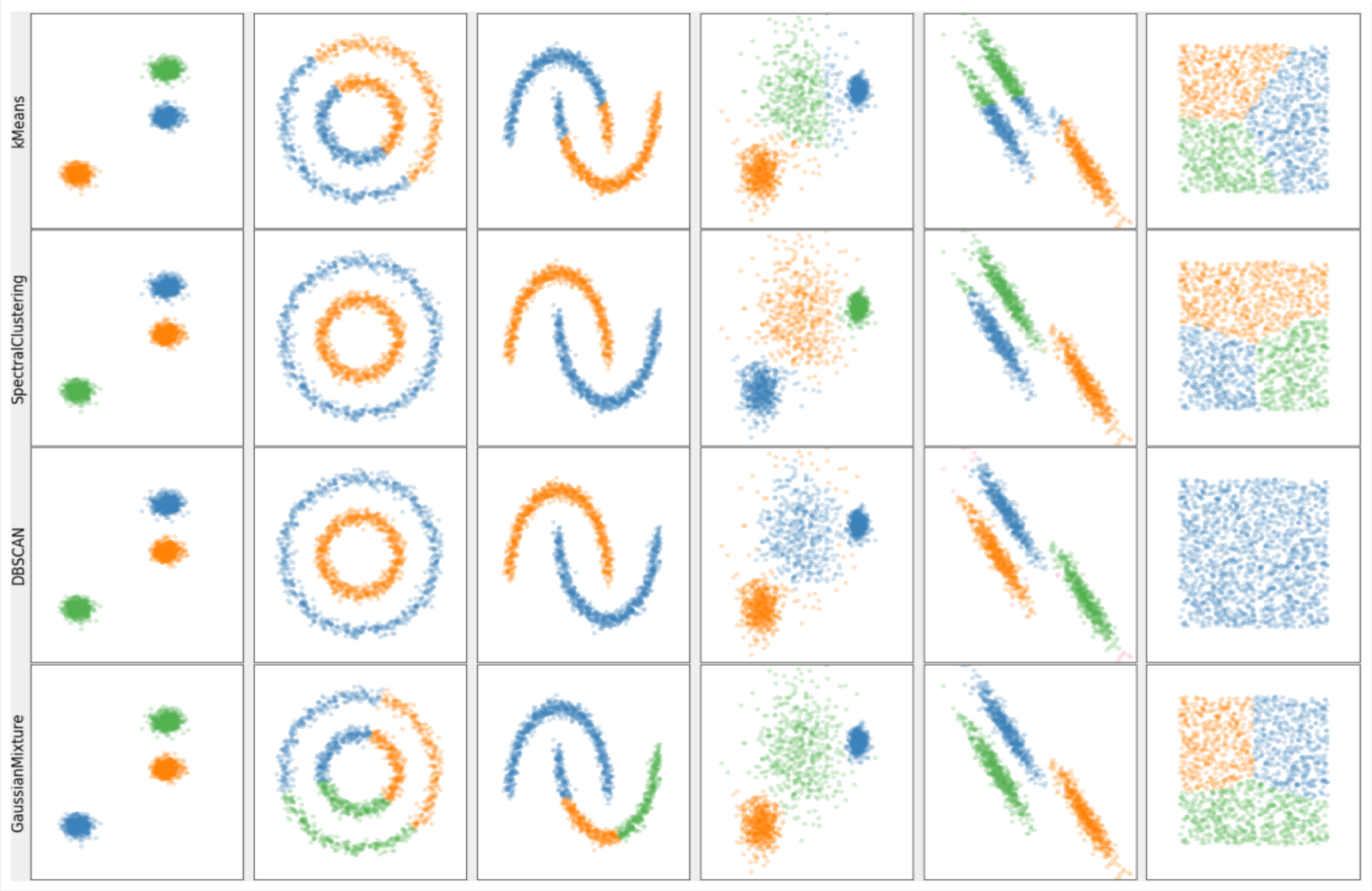
Choosing the value for k is not automatically done by the algorithm, the value of k is usually manually selected, or derived from some measure. The most common of these methods is the **elbow method** [@Thorndike 1943]:

Elbow Method



Geometry of the Dataset

Simple k-Means clustering is only good for specific geometries of the data. The model works best only on gaussian distributed blobs with even cluster sizes. For more complex geometries, more complex algorithms like, DBSCAN, spectral clustering or gaussian mixture can be used [@scikitlearn2017].


kMeans
SpectralClustering
DBSCAN
GaussianMixture