# In unsupervised learning, if a ground truth about a dataset is unknown, how can we determine the most useful number of clusters to be?

With **supervised** learning, the number of classes in a particular set of data is known outright, since each data instance in labeled as a member of a particular existent class. In the worst case, we can scan the class attribute and count up the number of unique entries which exist.

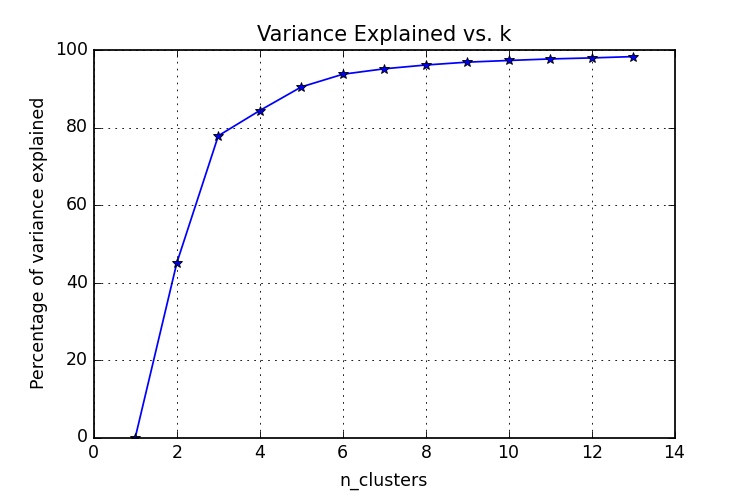
With [**unsupervised**](https://www.kdnuggets.com/tag/unsupervised-learning) learning, the idea of class attributes and explicit class membership does not exist; in fact, one of the dominant forms of unsupervised learning -- data clustering -- aims to approximate class membership by minimizing interclass instance similarity and maximizing intraclass similarity. A major drawback with clustering can be the requirement to provide the number of classes which exist in the unlabeled dataset at the onset, in some form or another. If we are lucky, we may know the data’s **ground truth** -- the actual number of classes -- beforehand. However, this is not always the case, for numerous reasons, one of which being that there may actually be no defined number of classes (and hence, clusters) in the data, with the whole point of the unsupervised learning task being to survey the data and attempt to impose some meaningful structure of optimal cluster and class numbers upon it.

Without knowing the ground truth of a dataset, then, how do we know what the optimal number of data clusters are? As one may expect, there are actually numerous methods to go about answering this question. We will have a look at 2 particular popular methods for attempting to answer this question: the elbow method and the silhouette method.

**The Elbow Method**

The elbow method is often the best place to state, and is especially useful due to its ease of explanation and verification via visualization. The elbow method is interested in explaining variance as a function of cluster numbers (the *k* in *k*-means). By plotting the percentage of variance explained against *k*, the first *N* clusters should add significant information, explaining variance; yet, some eventual value of *k* will result in a much less significant gain in information, and it is at this point that the graph will provide a noticeable angle. This angle will be the optimal number of clusters, from the perspective of the elbow method,

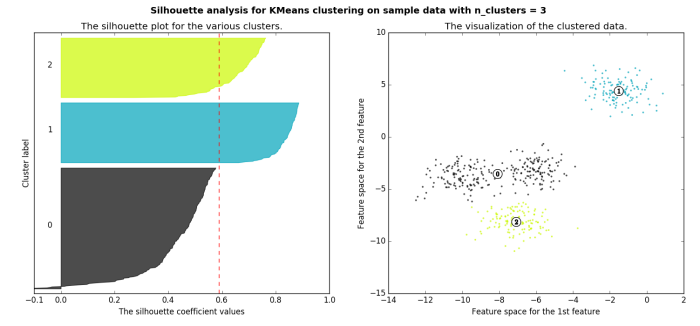
It should be self-evident that, in order to plot this variance against varying numbers of clusters, varying numbers of clusters must be tested. Successive complete iterations of the clustering method must be undertaken, after which the results can be plotted and compared.

  
[Image source](http://elf11.github.io/2015/08/18/Kmeans-analysis.html).

**The Silhouette Method**

The silhouette method measures the similarity of an object to its own cluster -- called cohesion -- when compared to other clusters -- called separation. The silhouette value is the means for this comparison, which is a value of the range [-1, 1]; a value close to 1 indicates a close relationship with objects in its own cluster, while a value close to -1 indicates the opposite.

A clustered set of data in a model producing mostly high silhouette values is likely an acceptable and appropriate model.

  
[Image source](http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html).