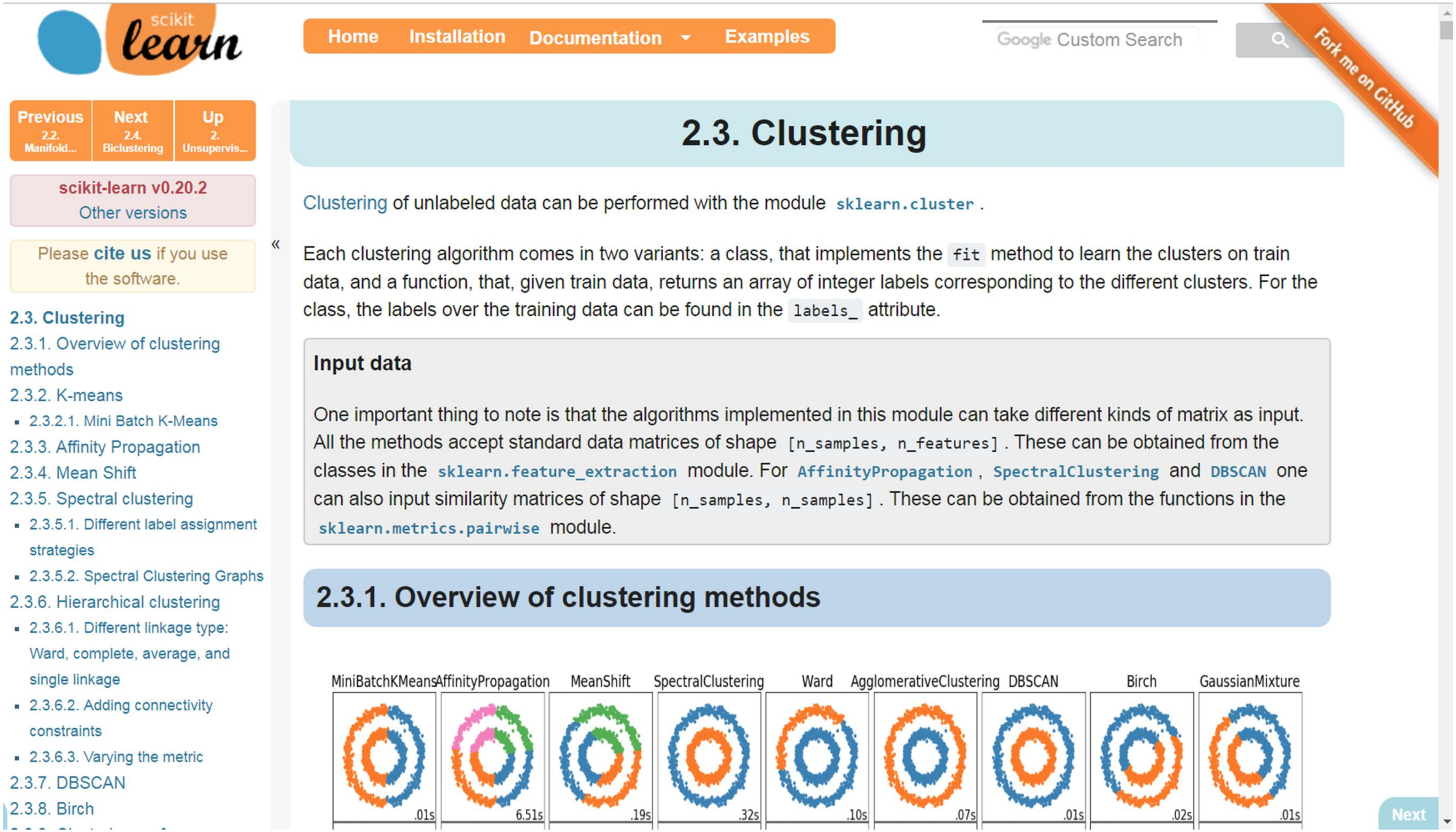


2.3. Clustering

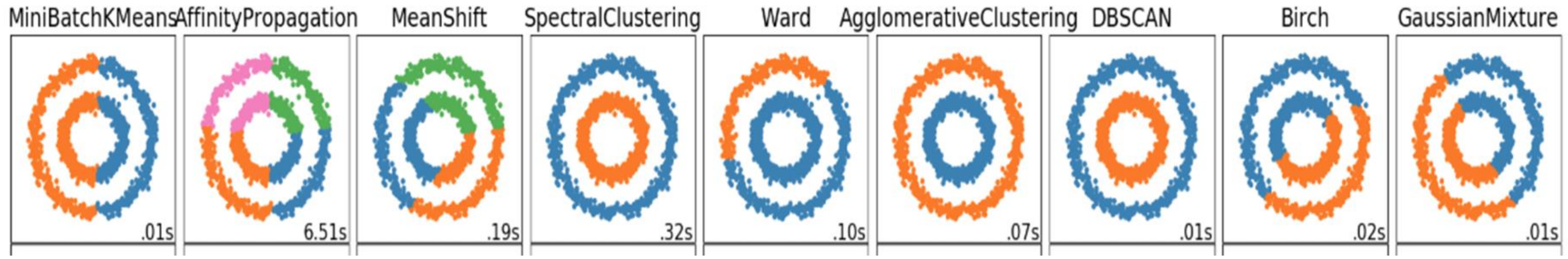
Clustering of unlabeled data can be performed with the module `sklearn.cluster`.

Each clustering algorithm comes in two variants: a class, that implements the `fit` method to learn the clusters on train data, and a function, that, given train data, returns an array of integer labels corresponding to the different clusters. For the class, the labels over the training data can be found in the `labels_` attribute.

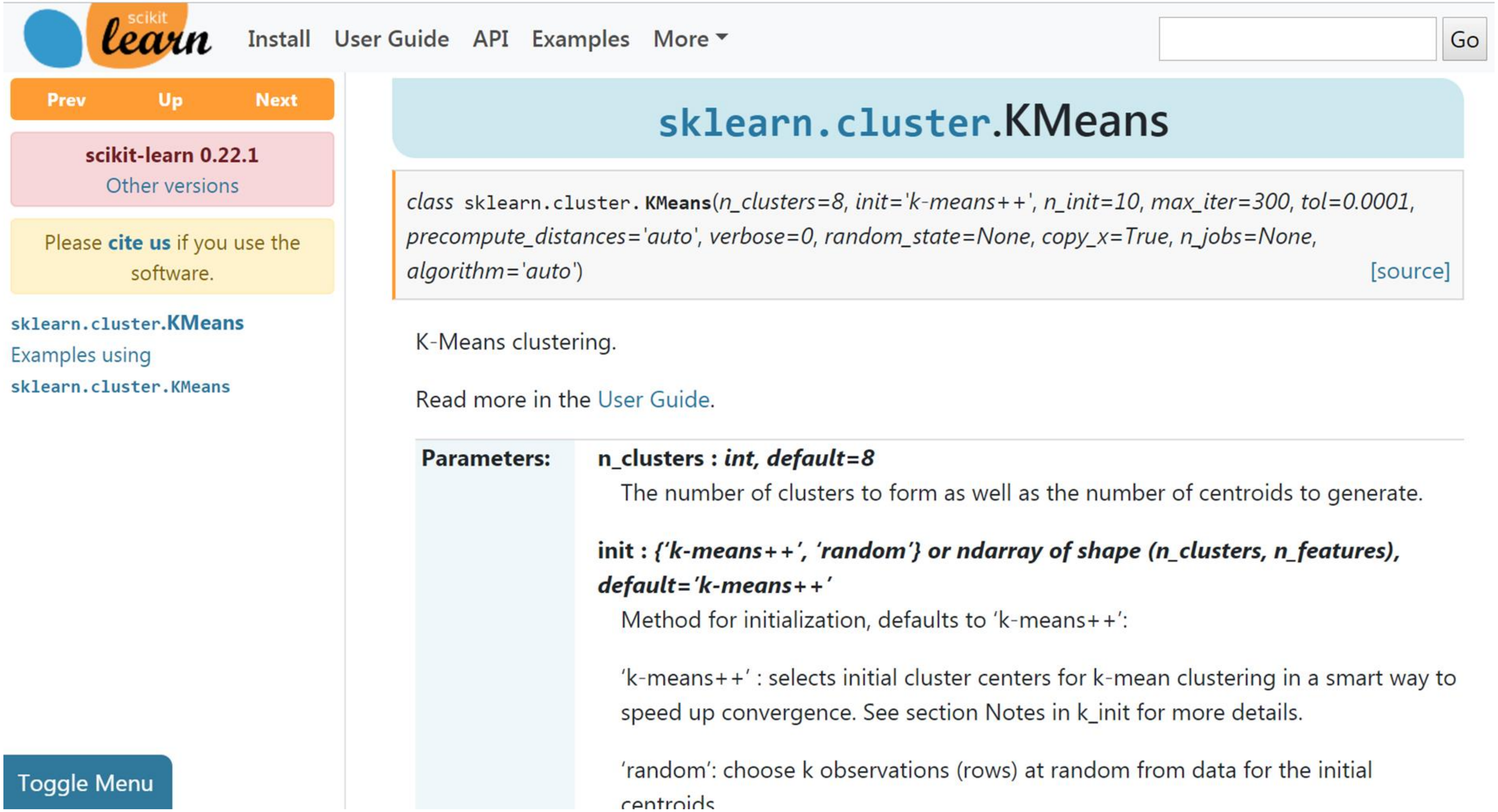
Input data

One important thing to note is that the algorithms implemented in this module can take different kinds of matrix as input. All the methods accept standard data matrices of shape `[n_samples, n_features]`. These can be obtained from the classes in the `sklearn.feature_extraction` module. For `AffinityPropagation`, `SpectralClustering` and `DBSCAN` one can also input similarity matrices of shape `[n_samples, n_samples]`. These can be obtained from the functions in the `sklearn.metrics.pairwise` module.

2.3.1. Overview of clustering methods



Next



Please **cite us** if you use the software.

`sklearn.cluster.`**KMeans**
Examples using `sklearn.cluster.KMeans`

`sklearn.cluster`.KMeans

class `sklearn.cluster.`**`KMeans`**(*n_clusters=8, init='k-means++', n_init=10, max_iter=300, tol=0.0001, precompute_distances='auto', verbose=0, random_state=None, copy_x=True, n_jobs=None, algorithm='auto'*) [source]

K-Means clustering.

Read more in the User Guide.

Parameters:

n_clusters : *int, default=8*

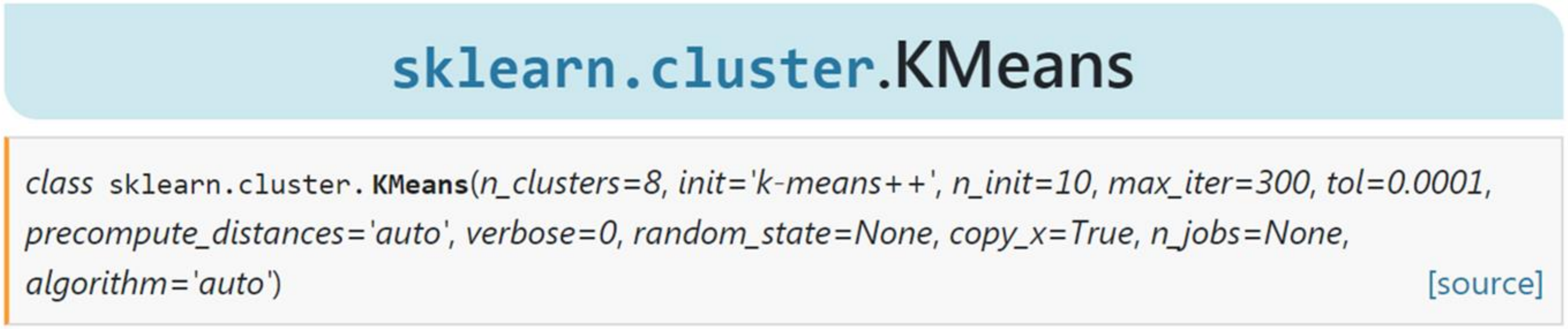
The number of clusters to form as well as the number of centroids to generate.

init : *{'k-means++', 'random'} or ndarray of shape (n_clusters, n_features), default='k-means++'*

Method for initialization, defaults to 'k-means++':

'k-means++' : selects initial cluster centers for k-mean clustering in a smart way to speed up convergence. See section Notes in k_init for more details.

'random': choose k observations (rows) at random from data for the initial centroids



sklearn.cluster.KMeans

```
class sklearn.cluster.KMeans(n_clusters=8, init='k-means++', n_init=10, max_iter=300, tol=0.0001,
precompute_distances='auto', verbose=0, random_state=None, copy_x=True, n_jobs=None,
algorithm='auto')
```

[source]

To use k-means, we need to understand the meaning of each parameter
The default parameter values may NOT work for your application.
You need to adjust the parameters

To understand the meanings of the parameters, we need to understand the algorithm of k-means

Run kmeans_cust_seg.ipynb