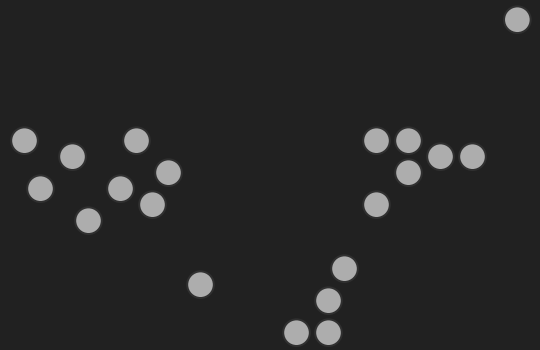
Clustering

Given a cloud of unlabelled points, identify consistent clusters



Clustering

Given a cloud of unlabelled points, identify consistent clusters

Why is it challenging?

- Small vs high dimension
- Small vs large datasets
- Ground truth often not available
- Scale matters

Examples of applications

Online products, user groups, image regions, market segments, ...

Astronomical objects from the Sloan Digital Sky Survey

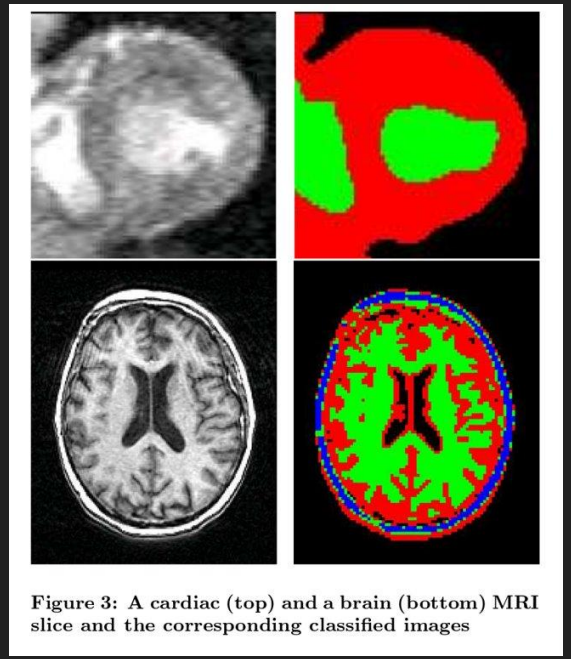
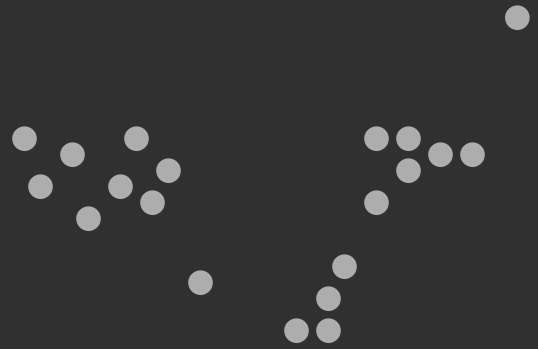


Figure 3: A cardiac (top) and a brain (bottom) MRI slice and the corresponding classified images

kmeans clustering

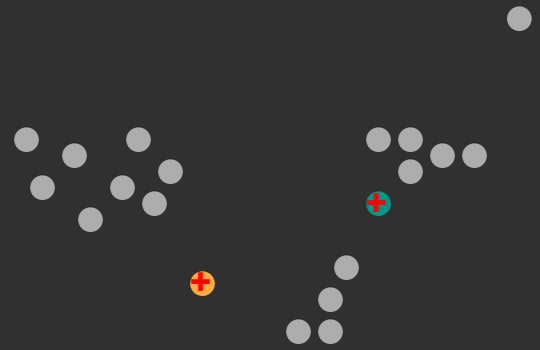
1. Pick k

here k=2



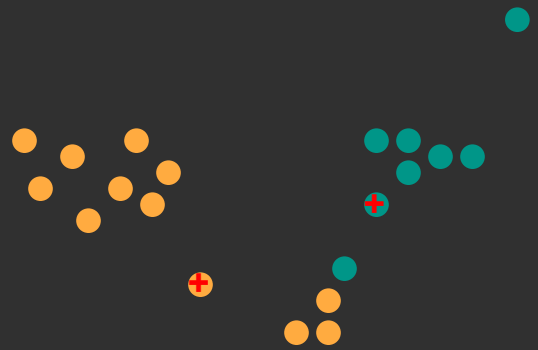
kmeans clustering

2. Initialize centroids randomly



kmeans clustering

3. Assign data points to closest cluster

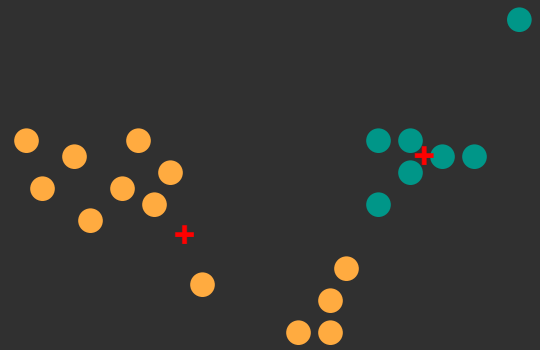


kmeans clustering

4. Update centroids

kmeans clustering

5. Repeat until convergence



kmeans clustering

5. Repeat until convergence

kmeans clustering

5. Repeat until convergence

> Converged

Pros and cons

Simple

Efficient, O(tkn)

Easy to parallelize

Guaranteed convergence

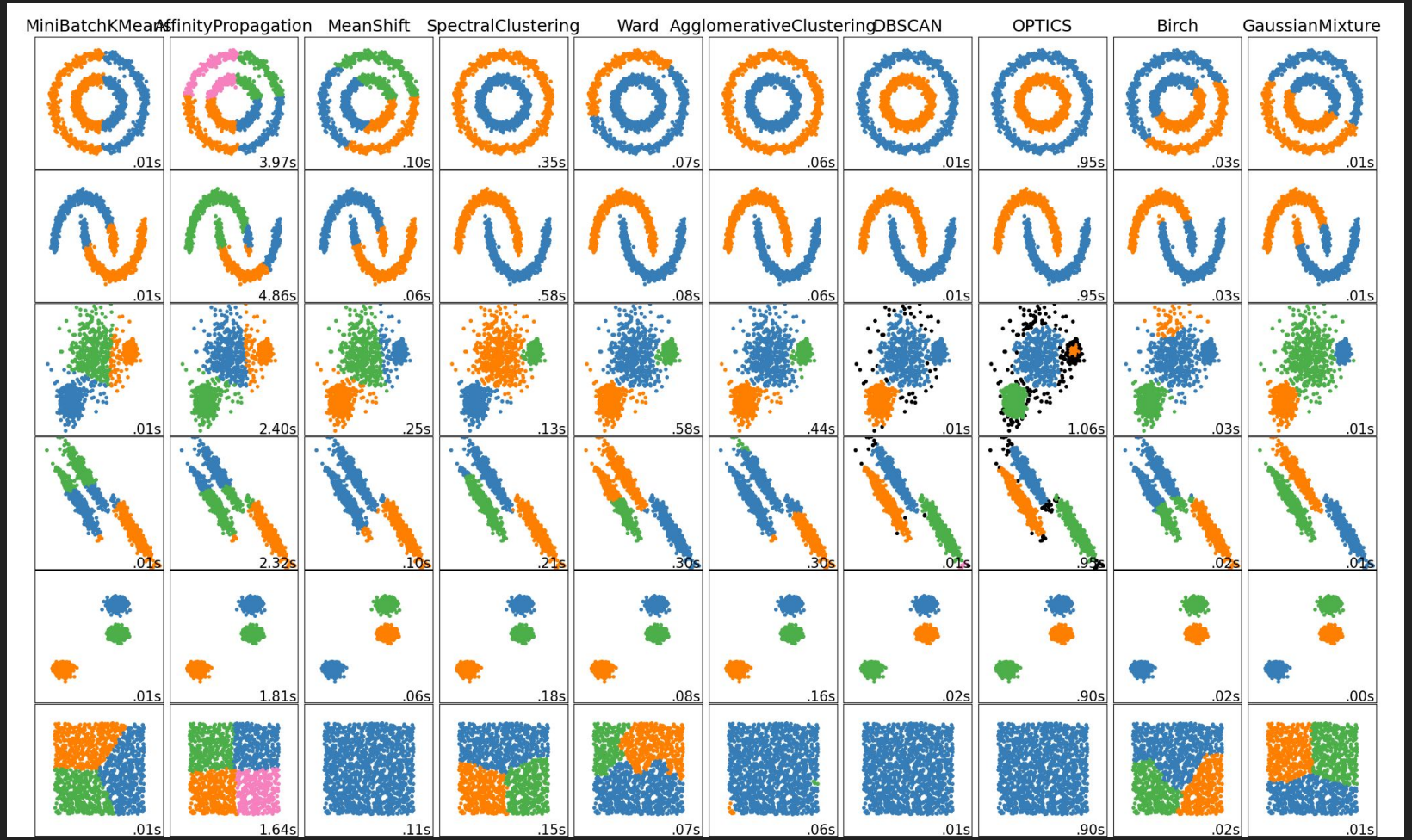
Need to set k

Mean must be defined

Sensitive to outliers

Sensitive to initialization

Assumes normally distributed clusters



A comparison of the clustering algorithms in [scikit-learn]