# INTRO TO DATA SCIENCE

# CLUSTERING & K-MEANS CLUSTERING

## OUTLINE

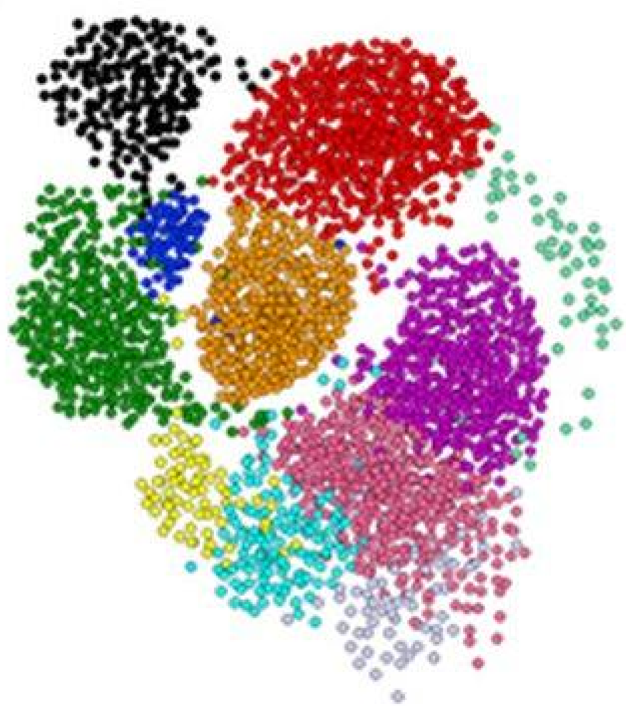
- **CLUSTERING**
- **K-MEANS CLUSTERING**
- **SELECTING K**

- **DEMO:**
  **K-MEANS CLUSTERING with SKLEARN**

# CLUSTERING

# CLUSTERING

- **Clustering**, or **cluster analysis**, is the task of grouping observations such that members of the same group, or **cluster**, are more similar to each other by some metric than they are to the members of the other clusters



## QUESTIONS--

- Is there some underlying structure in the data?
  - unsupervised task, not predicting anything
- Do any sub-populations exist in the data?
  - how many are there? how big are they?
  - what are their common properties?
  - are there outliers?

## TYPES OF CLUSTERING METHODS

- Hard clustering:
  - clusters do not overlap-- item belongs to a single cluster
- Soft clustering:
  - clusters can overlap-- probability of membership in a cluster

## CLUSTER ANALYSIS

Q: What is a cluster?

Q: What is a cluster?

A: A group of ***similar*** data points

Q: What is a cluster?

A: A group of ***similar*** data points

The concept of ***similarity*** is central to the definition of a cluster, and therefore to cluster analysis

Q: How do you solve a clustering problem?

## CLUSTER ANALYSIS

Q: How do you solve a clustering problem?

A: Think of a cluster as a “potential class”; then the solution to a clustering problem is to programmatically determine these classes

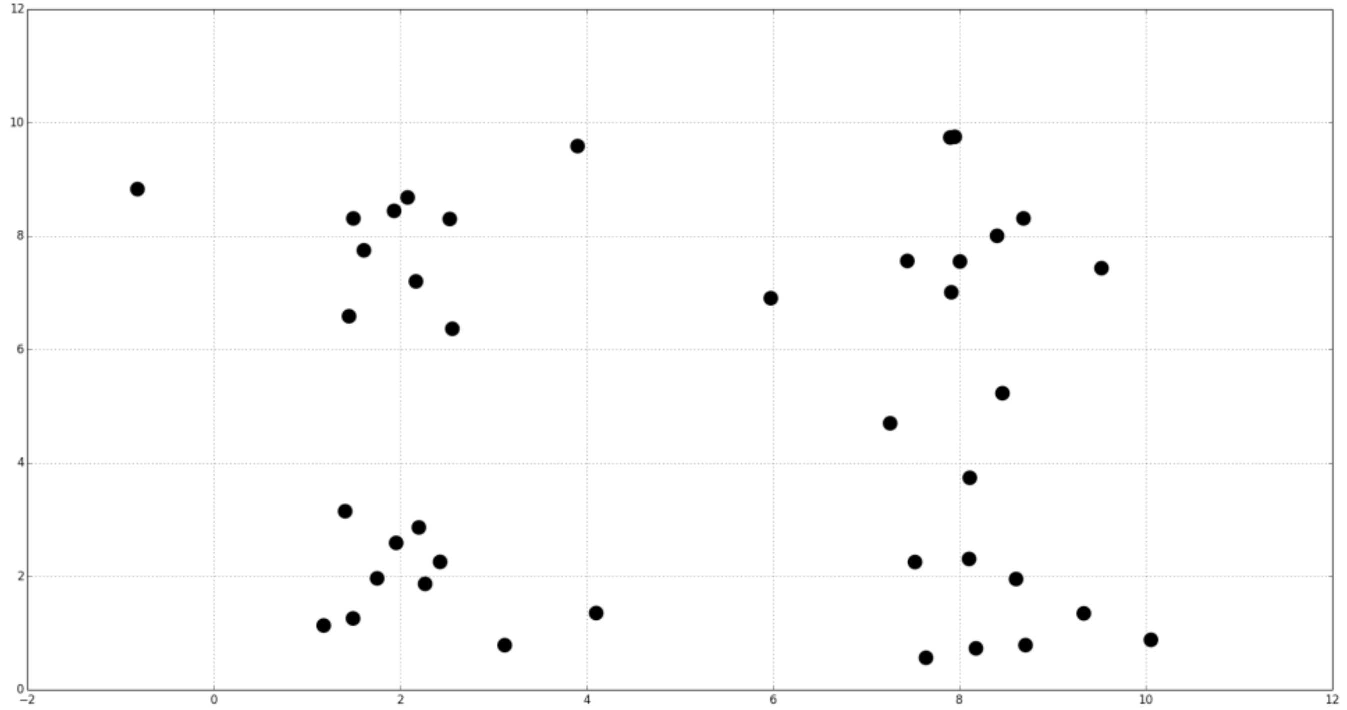
## CLUSTER ANALYSIS

Q: How do you solve a clustering problem?

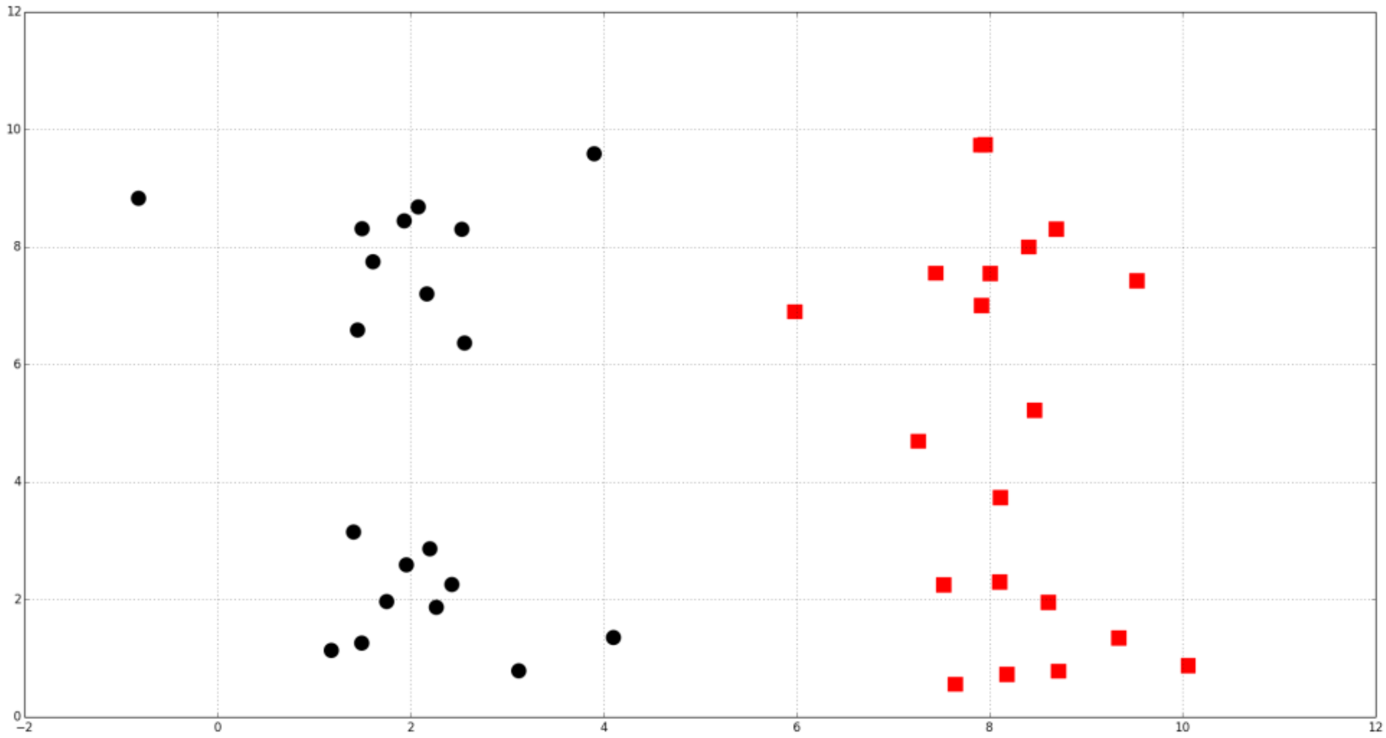
A: Think of a cluster as a “potential class”; then the solution to a clustering problem is to programmatically determine these classes

A goal of clustering can be data exploration, so a solution is anything that contributes to your understanding

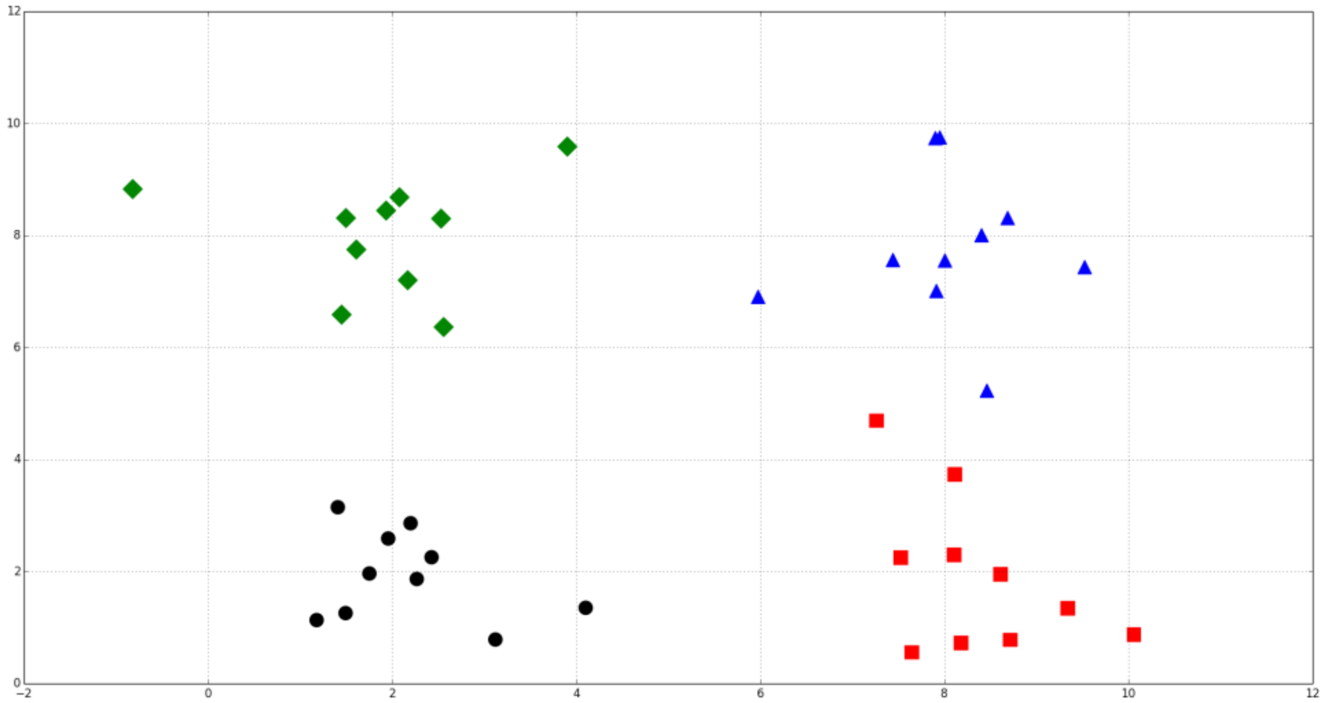
# CLUSTERING



# CLUSTERING



# CLUSTERING



## APPLICATIONS OF CLUSTERING

- Data exploration
- 

## APPLICATIONS OF CLUSTERING

- Data exploration
- Identify communities, connections in social networks
- Customer segmentation
- Find groups of genes with similar expression patterns
- Recommendation systems
- Image compression

# K-means Clustering

# THE BASIC K-MEANS ALGORITHM

from wikipedia:

*“a method that is popular for cluster analysis in data mining. k-means clustering aims to partition **n** observations into **k** clusters in which each observation belongs to the cluster with the nearest mean”*

## THE BASIC K-MEANS ALGORITHM

1. choose ***k*** initial centroids (note that k is an input)

## THE BASIC K-MEANS ALGORITHM

1. choose ***k*** initial centroids (note that k is an input)
2. for each data point:
   - find distance to each centroid (k)
   - assign point to nearest centroid

## THE BASIC K-MEANS ALGORITHM

1. choose ***k*** initial centroids (note that k is an input)
2. for each data point:
   - find distance to each centroid (k)
   - assign point to nearest centroid
3. recalculate centroid positions

## THE BASIC K-MEANS ALGORITHM

1. choose ***k*** initial centroids (note that k is an input)
2. for each data point:
   - find distance to each centroid (k)
   - assign point to nearest centroid
3. recalculate centroid positions
4. repeat steps 2-3 until stopping criteria met

# Demo

Visualizing K-means

## STEP 1 – CHOOSING INITIAL CENTROIDS

Q: How do you choose the initial centroid positions?

## STEP 1 – CHOOSING INITIAL CENTROIDS

Q: How do you choose the initial centroid positions?

A: There are several options, including:

## STEP 1 – CHOOSING INITIAL CENTROIDS

Q: How do you choose the initial centroid positions?

A: There are several options, including:

- randomly (but may yield divergent behavior)
- run alternative clustering task, use resulting centroids as initial k-means centroids

## STEP 2 – SIMILARITY MEASURES

Q: How do you determine which centroid is the nearest?

## STEP 2 – SIMILARITY MEASURES

Q: How do you determine which centroid is the nearest?

The “nearness” criterion is determined by a similarity/distance measure

## STEP 2 – SIMILARITY MEASURES

There are a number of different similarity measures to choose from, and in general the right choice depends on the problem.

## STEP 2 – SIMILARITY MEASURES

There are a number of different similarity measures to choose from, and in general the right choice depends on the problem.

For many datasets, the typical choice is the Euclidean distance:

$$d(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

## STEP 2 – SIMILARITY MEASURES

Ex: One popular metric for text mining problems (or any problem with *sparse binary* data) is the ***Jaccard*** coefficient,

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

## STEP 2 – SIMILARITY MEASURES

Ex: One popular metric for text mining problems (or any problem with *sparse binary* data) is the ***Jaccard*** coefficient,

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Applying this metric to a problem expresses the sparse nature of the data, and makes a variety of text mining techniques accessible.

## STEP 3 – OBJECTIVE FUNCTION

Q: How do we recompute the positions of the centroids at each iteration of the algorithm?

## STEP 3 – OBJECTIVE FUNCTION

Q: How do we recompute the positions of the centroids at each iteration of the algorithm?

A: By optimizing an objective function that tells us how “good” the clustering is.

## STEP 3 – OBJECTIVE FUNCTION

Q: How do we recompute the positions of the centroids at each iteration of the algorithm?

A: By optimizing an objective function that tells us how “good” the clustering is.

The iterative part of the algorithm (recomputing centroids and reassigning points to clusters) explicitly tries to minimize this objective function.

## STEP 3 – OBJECTIVE FUNCTION

Ex: Using the Euclidean distance measure, one typical objective function is the Sum of Squared Errors (SSE) from each point $\boldsymbol{x}$ to its centroid $\boldsymbol{c_i}$:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} d(x, c_i)^2$$

## STEP 3 – OBJECTIVE FUNCTION

Ex: Using the Euclidean distance measure, one typical objective function is the Sum of Squared Errors (SSE) from each point ***x*** to its centroid $\boldsymbol{c_i}$:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} d(x, c_i)^2$$

Given two clusterings, we will prefer the one with the lower SSE since this means the centroids have converged to better locations

## STEP 4 – CONVERGENCE

We iterate until some stopping criteria are met; in general, suitable convergence is achieved in a small number of steps.

## STEP 4 – CONVERGENCE

We iterate until some stopping criteria are met; in general, suitable convergence is achieved in a small number of steps.

Stopping criteria can be based on the centroids (eg, if centroid positions change by no more than $\varepsilon$) or on the points (eg, if no more than $x$% change clusters between iterations).

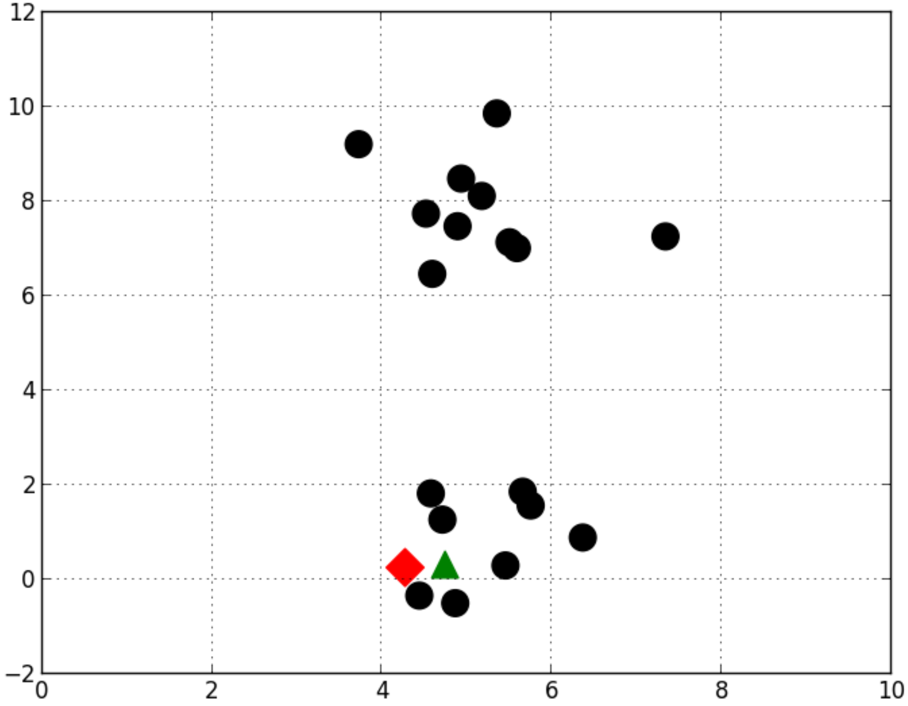
# ADVANTAGES OF K-MEANS

- K-Means is fast!
- Can be scaled to large data sets when using mini-batches
- Excellent for general-purpose clustering

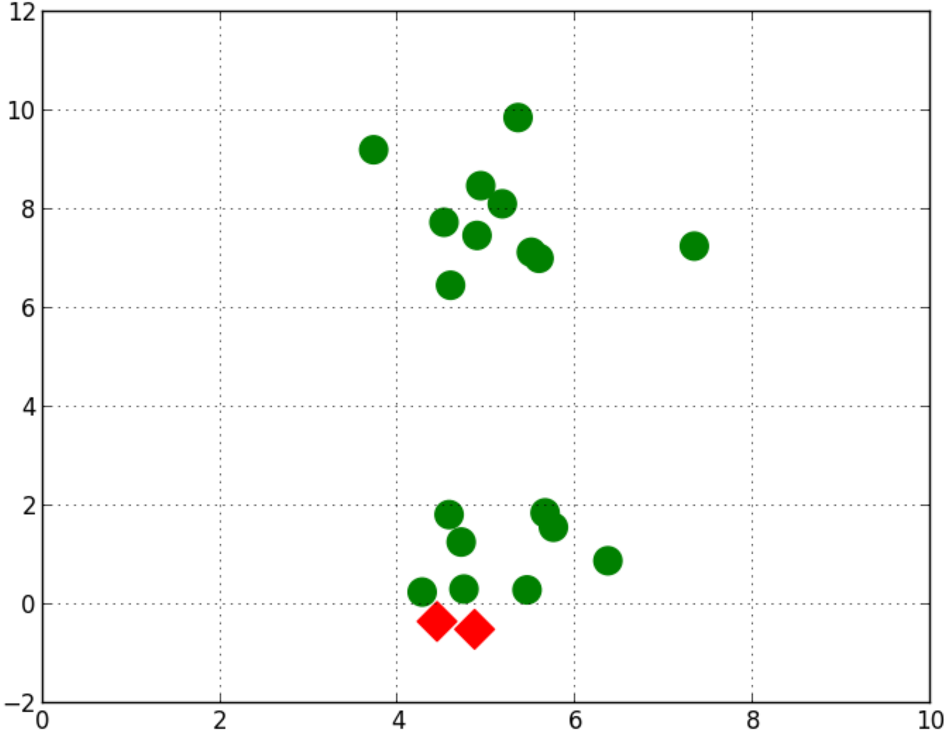
# DISADVANTAGES OF K-MEANS

- Random initializations can result in converging to ***local minima***
- Different random starting centroids can yield different results
- Nearby points can sometimes end up in different clusters
- Can be difficult to choose the right value for ***k***

# DISADVANTAGES OF K-MEANS

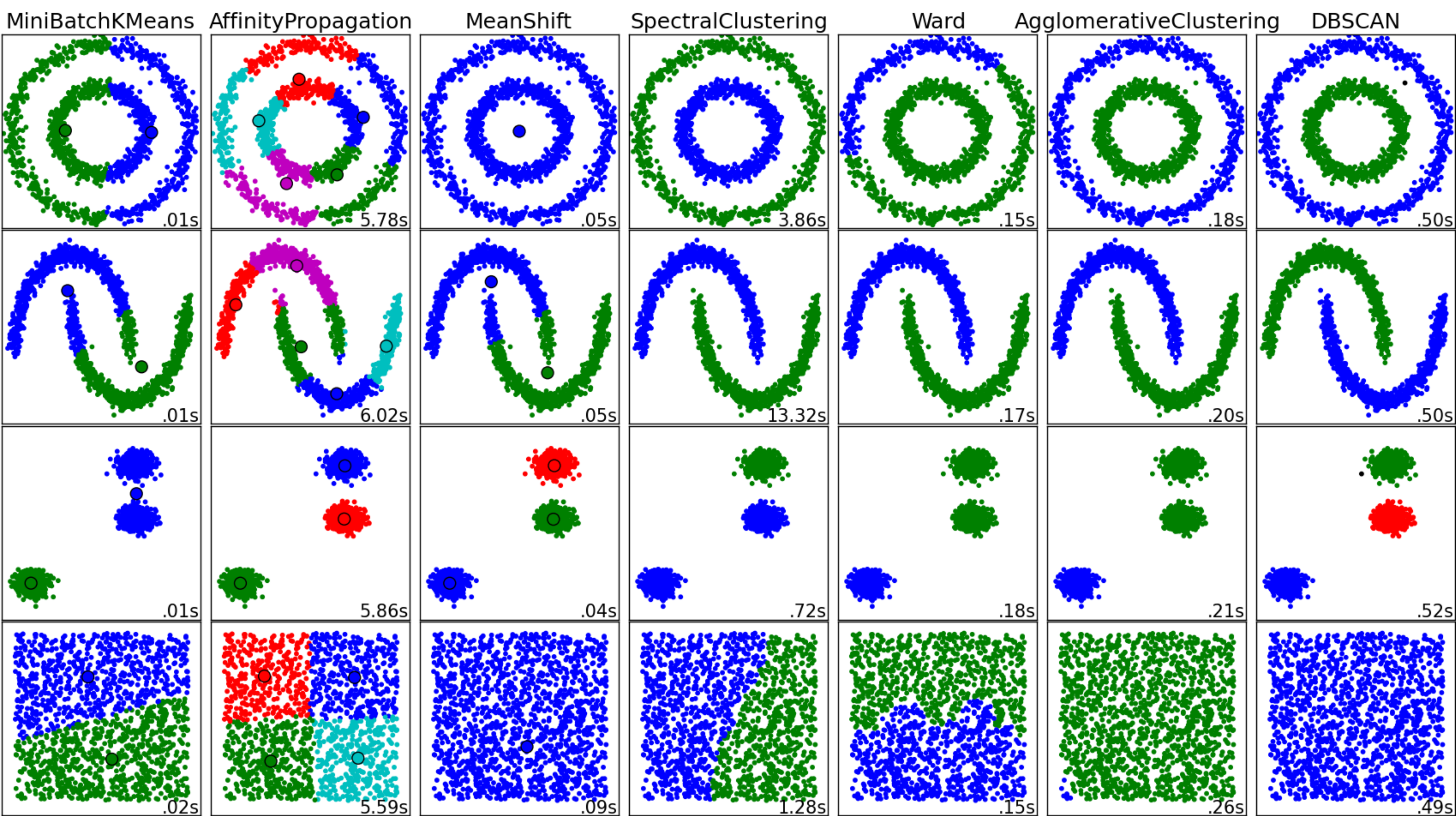


# DISADVANTAGES OF K-MEANS



# OTHER CLUSTERING ALGORITHMS

- Affinity Propagation
- MeanShift
- Spectral
- Ward
- Agglomerative
- DBSCAN


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
.01s
5.78s
.05s
3.86s
.15s
.18s
.50s
.01s
6.02s
.05s
13.32s
.17s
.20s
.50s
.01s
5.86s
.04s
.72s
.18s
.21s
.52s
.02s
5.59s
.09s
1.28s
.15s
.26s
.49s

# SELECTING *K WITH THE ELBOW METHOD*

## SELECTING K WITH THE ELBOW METHOD

- The elbow method plots the value of the cost function produced by different values of ***k***

## SELECTING K WITH THE ELBOW METHOD

- The elbow method plots the value of the cost function produced by different values of ***k***

- As ***k*** increases, the average distortion will decrease; each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids

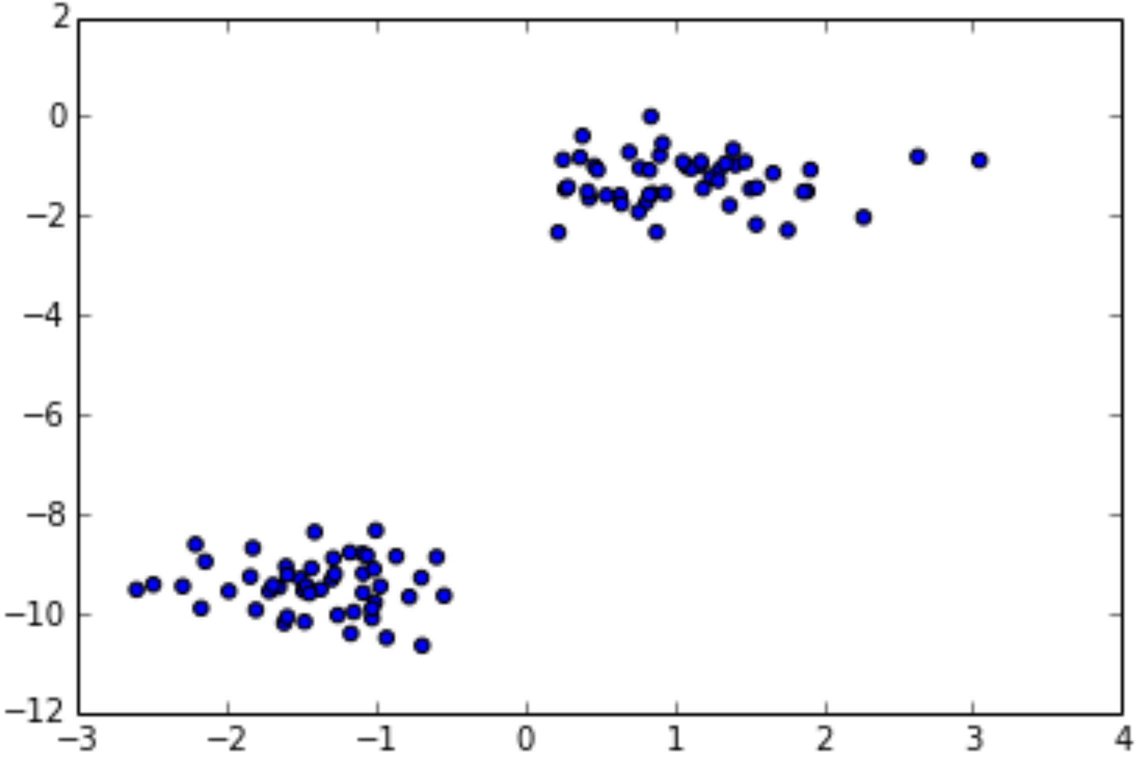
## SELECTING K WITH THE ELBOW METHOD

- The elbow method plots the value of the cost function produced by different values of ***k***

- As ***k*** increases, the average distortion will decrease; each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids

- However, the improvements to the average dispersion will decline as ***k*** increases. The value of ***k*** at which the improvement to the dispersion declines the most is called the elbow

# SELECTING K WITH THE ELBOW METHOD



# SELECTING K WITH THE ELBOW METHOD