CS 405/605 Data Science

Dr. Qianqian Tong

Course Schedule

Week 13: April 3 topic: Random Forest
April 5 topic: Validation
Project stage IV and V released, and the ddl will be April 28.

Week 14: April 10 topic: PCA
April 12 topic: Clustering-Kmeans

Week 15: April 17 topic: Visualization
April 19 topic: Visualization
HW3 will be released, and the ddl will be April 28.

Week 16: April 24   Project Presentation (4 groups, each will have 15-20 min)
April 26   Project Presentation (4 groups, each will have 15-20 min)
All reports and homework must be submitted by April 28, and graded by the final week.

UNC
GREENSBORO

K-means Clustering

Clustering:

- What is clustering?
- Types of clustering
- Clustering algorithm ----- k-means
- How can you do this efficiently?

Clustering:

- What is clustering?

K-means Clustering

Clustering:

Task 1 : Group These Set of Document into 3 Groups based on meaning

Doc1 : Health , Medicine, Doctor

Doc 2 : Machine Learning, Computer

Doc 3 : Environment, Planet

Doc 4 : Pollution, Climate Crisis

Doc 5 : Covid, Health , Doctor

UNC GREENSBORO

K-means Clustering

Clustering:

Task 1 : Group These Set of Document into 3 Groups.

Doc1 : Health , Medicine, Doctor

Doc 2 : Machine Learning, Computer

Doc 3 : Environment, Planet
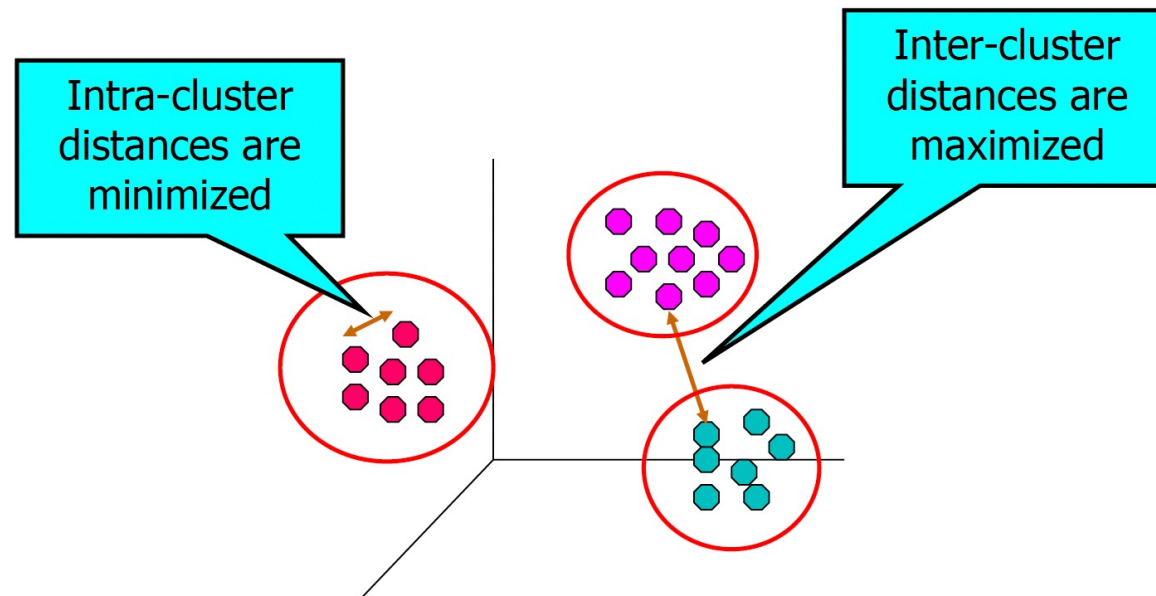
Doc 4 : Pollution, Climate Crisis

Doc 5 : Covid, Health , Doctor

K-means Clustering

Clustering:

Task 1 : Group These Set of Document into 3 Groups.

Doc 3 : Environment, Planet

Doc1 : Health , Medicine, Doctor

Doc 5 : Covid, Health , Doctor

Doc 4 : Pollution, Climate Crisis

Doc 2 : Machine Learning, Computer

UNC
GREENSBORO

Clustering:

- What is clustering?

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

Intra-cluster distances are minimized

Inter-cluster distances are maximized

UNC GREENSBORO

K-means Clustering

Clustering:

- What is clustering?

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

Clustering:

- Types of clustering

  o A *clustering* is a set of clusters
  o Important distinction between *hierarchical* and *partitional* sets of clusters
  o *Hierarchical clustering*
     – *A set of nested clusters organized as a hierarchical Tree*
  o *Partitional Clustering*
     – *A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset*
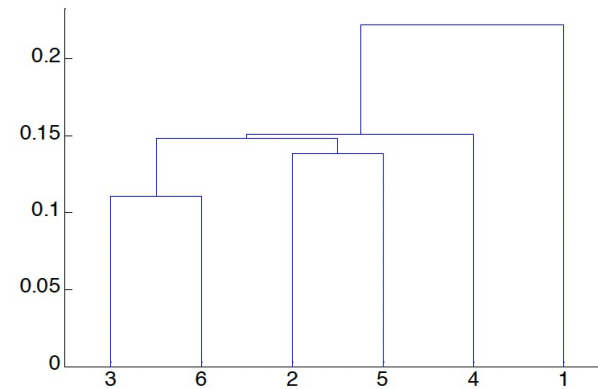
Clustering:

- Types of clustering

Hierarchical clustering
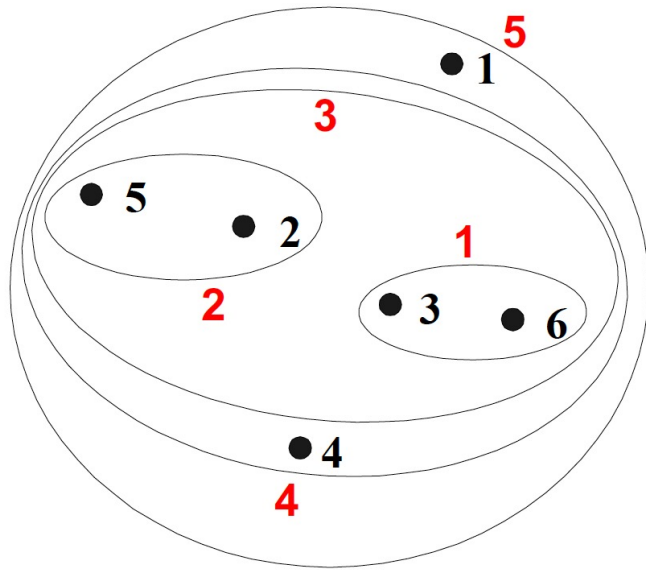
Nested Clusters   Dendrogram

Clustering:

- Types of clustering

Partitional Clustering
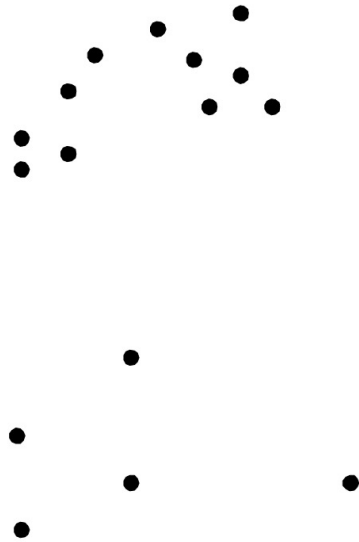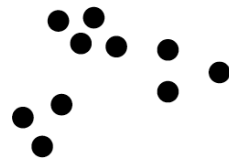

Original Points
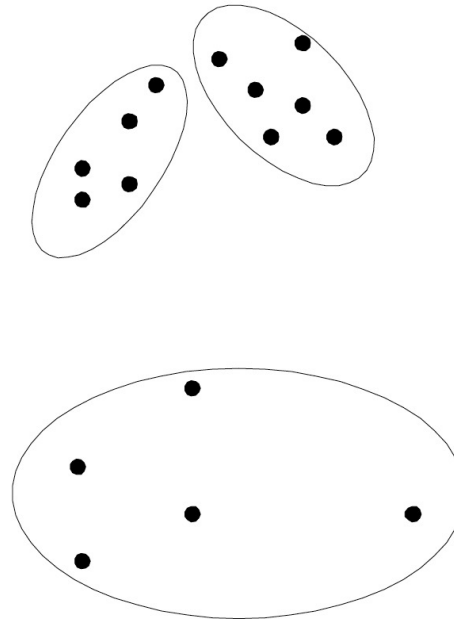
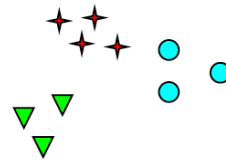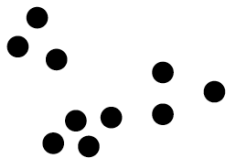**A Partitional Clustering**
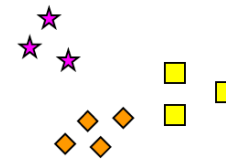
K-means Clustering

Clustering:

- Types of clustering
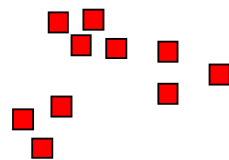
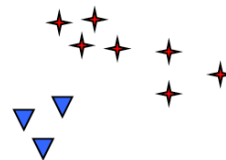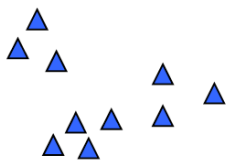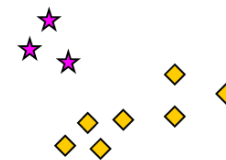*Partitional Clustering*

How many clusters?

Six Clusters

Notion of a Cluster can be Ambiguous!

Two Clusters

Four Clusters

Clustering:

- *Clustering Algorithms*

  - *K-means and its variants*
  - *Hierarchical clustering*
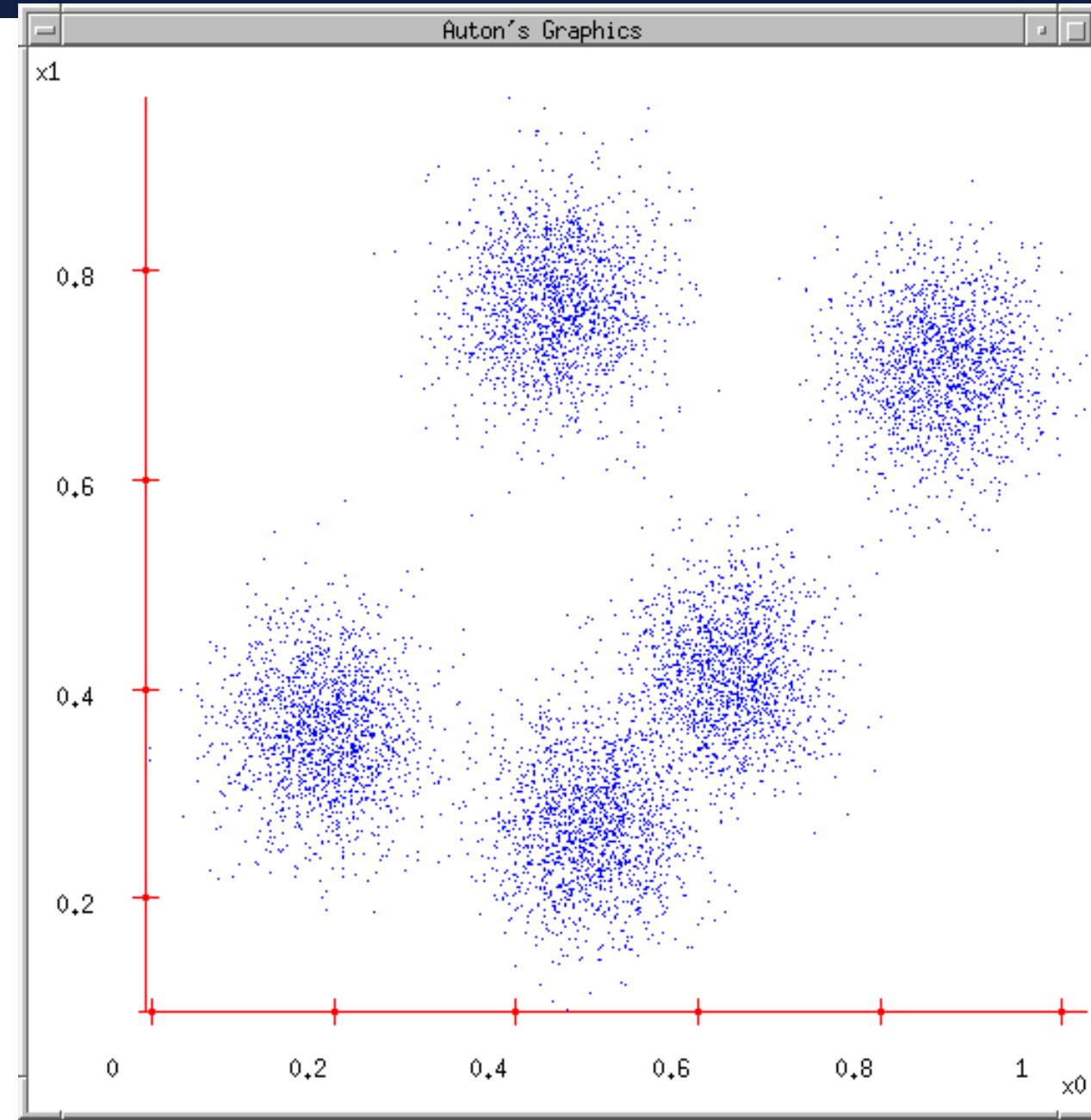 - *Density-based clustering*
  - *Spectral clustering*

K-means Clustering

Clustering:

- *Clustering Algorithms*

  - *K-means*

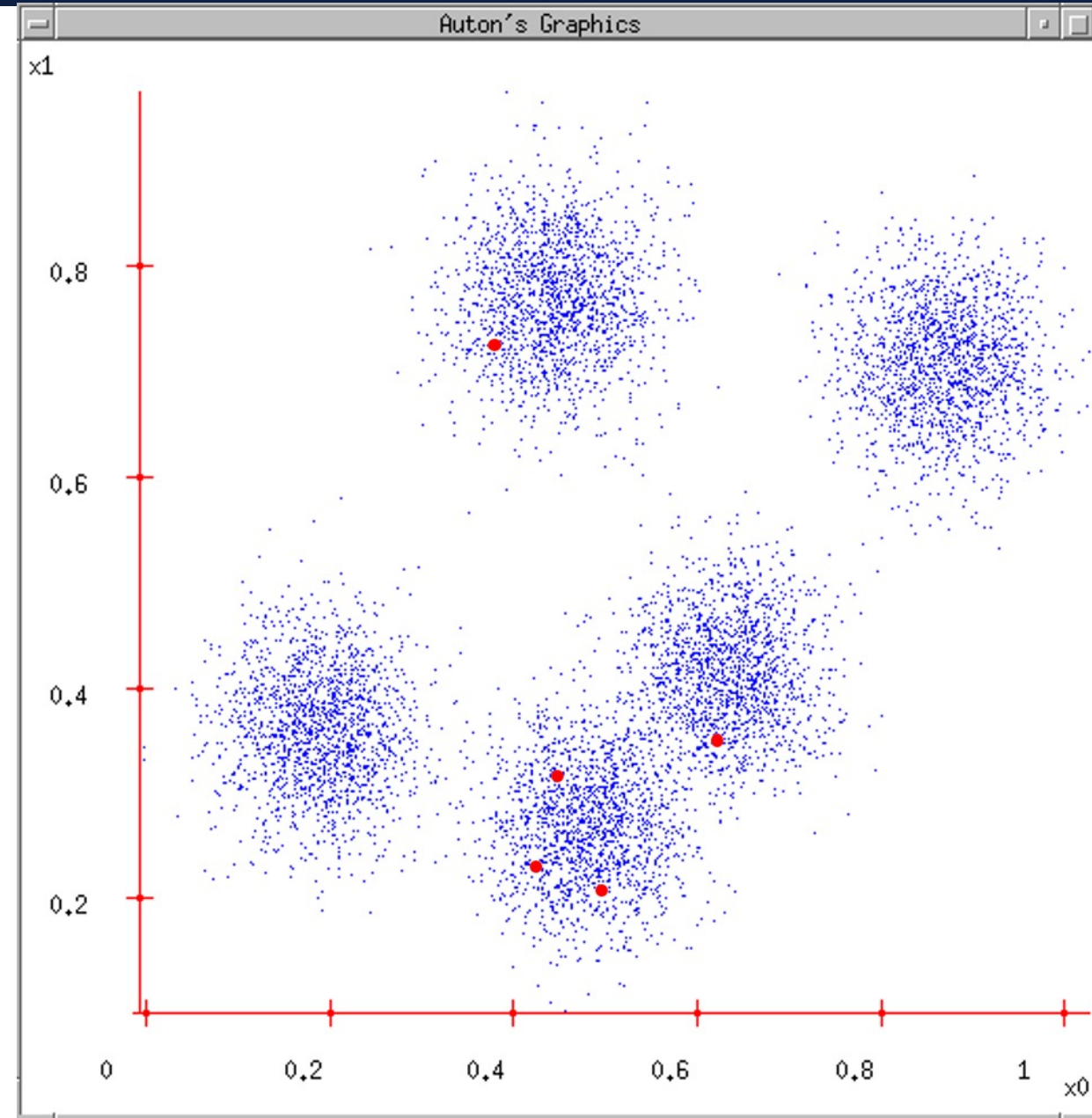1.Ask user how many clusters they'd like. *(e.g. k=5)*

K-means Clustering

Clustering:

- *Clustering Algorithms*

  - *K-means*

1. Ask user how many clusters they'd like. *(e.g. k=5)*

2. Randomly guess k cluster Center locations

K-means Clustering

Clustering:

- *Clustering Algorithms*

  - *K-means*

1.Ask user how many clusters they'd like. *(e.g. k=5)*

2.Randomly guess k cluster Center locations

3.Each datapoint finds out which Center it's closest to.

  (Thus each Center "owns" a set of datapoints)

Clustering:

- *Clustering Algorithms*

  - *K-means*

1. Ask user how many clusters they'd like. *(e.g. k=5)*

2. Randomly guess k cluster Center locations

3. Each datapoint finds out which Center it's closest to.

4. Each Center finds the centroid of the points it owns

K-means Clustering

Clustering:

- *Clustering Algorithms*

  - *K-means*

1. Ask user how many clusters they'd like. *(e.g. k=5)*
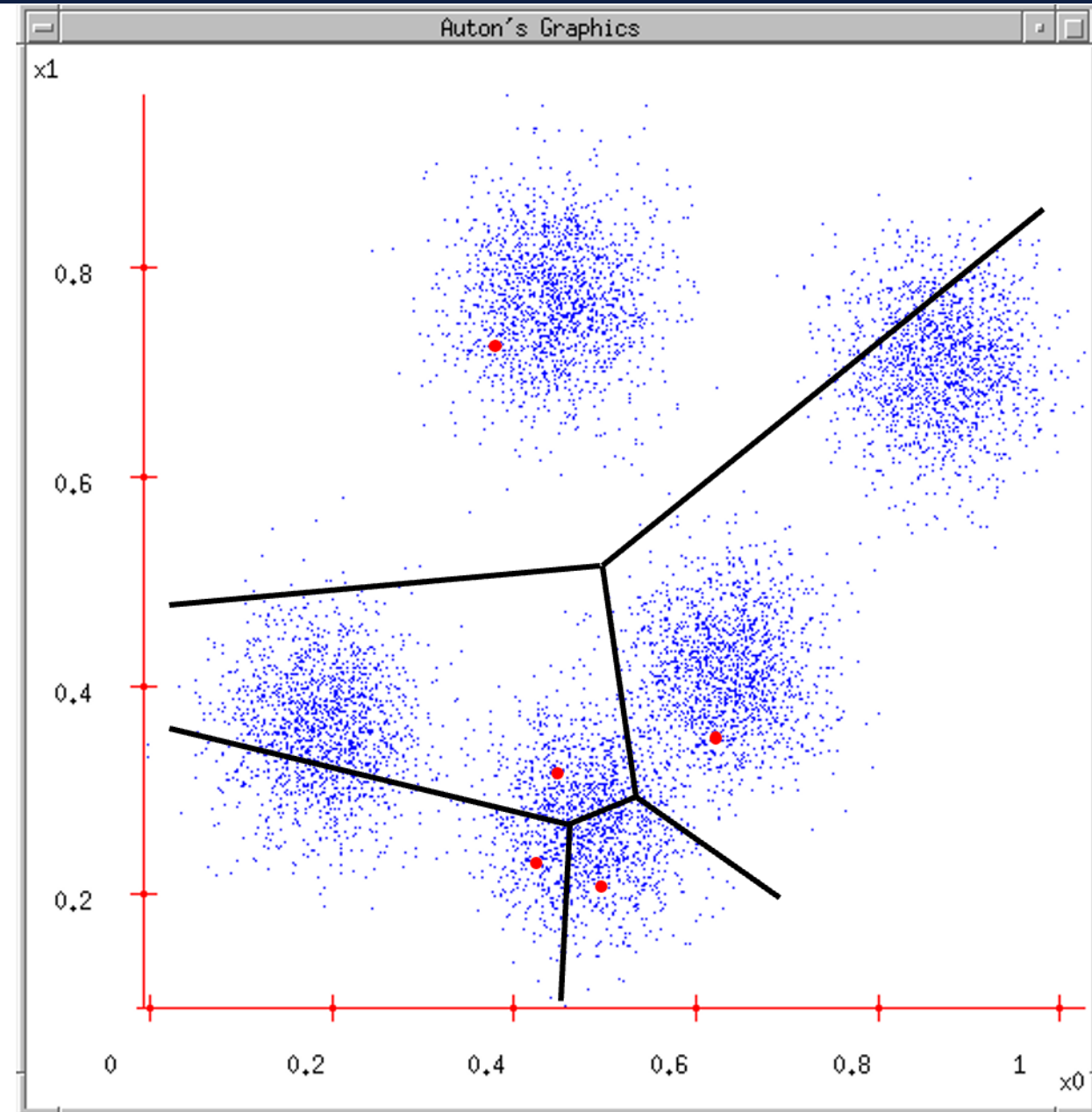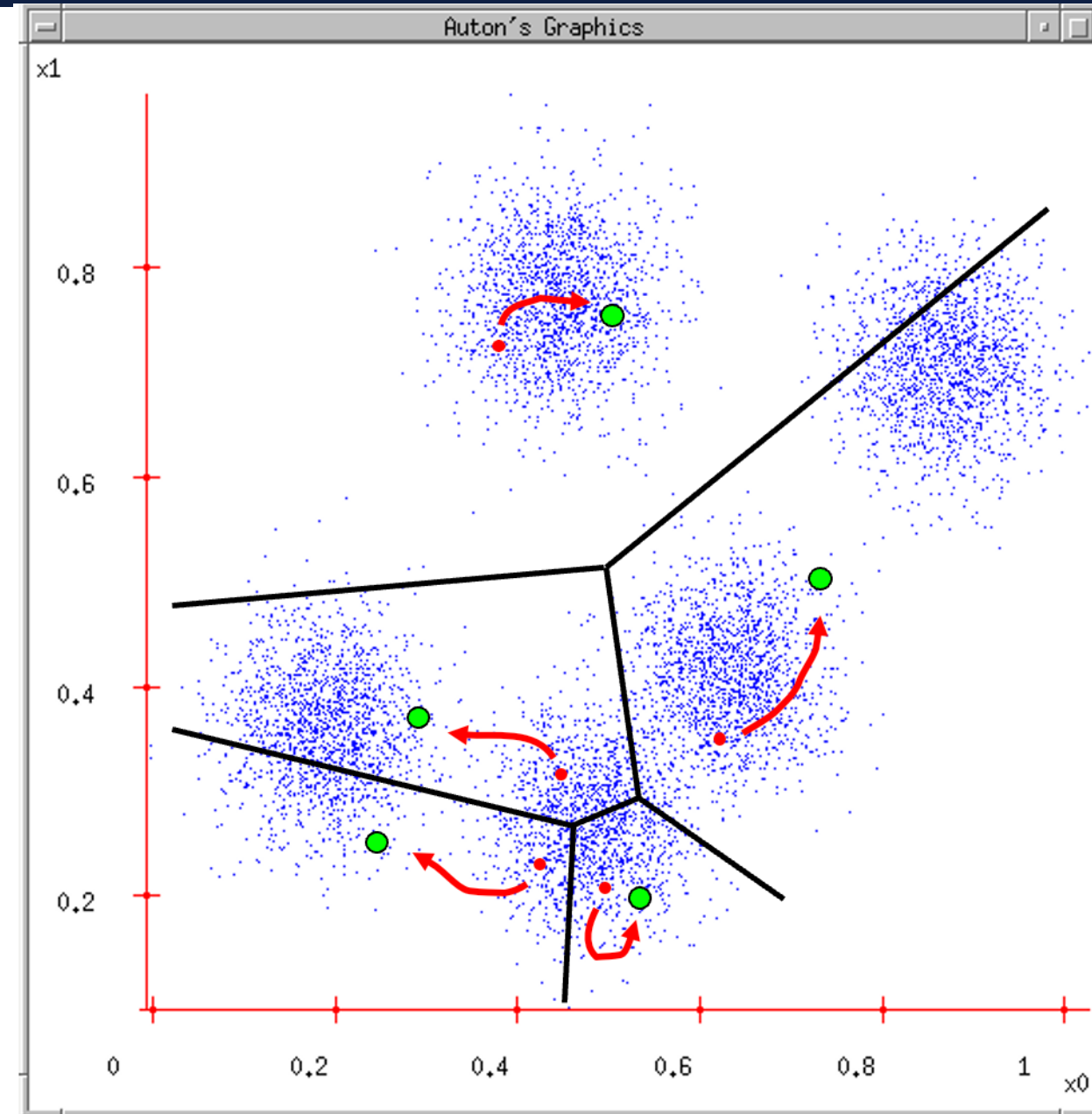
2. Randomly guess k cluster Center locations

3. Each datapoint finds out which Center it's closest to.

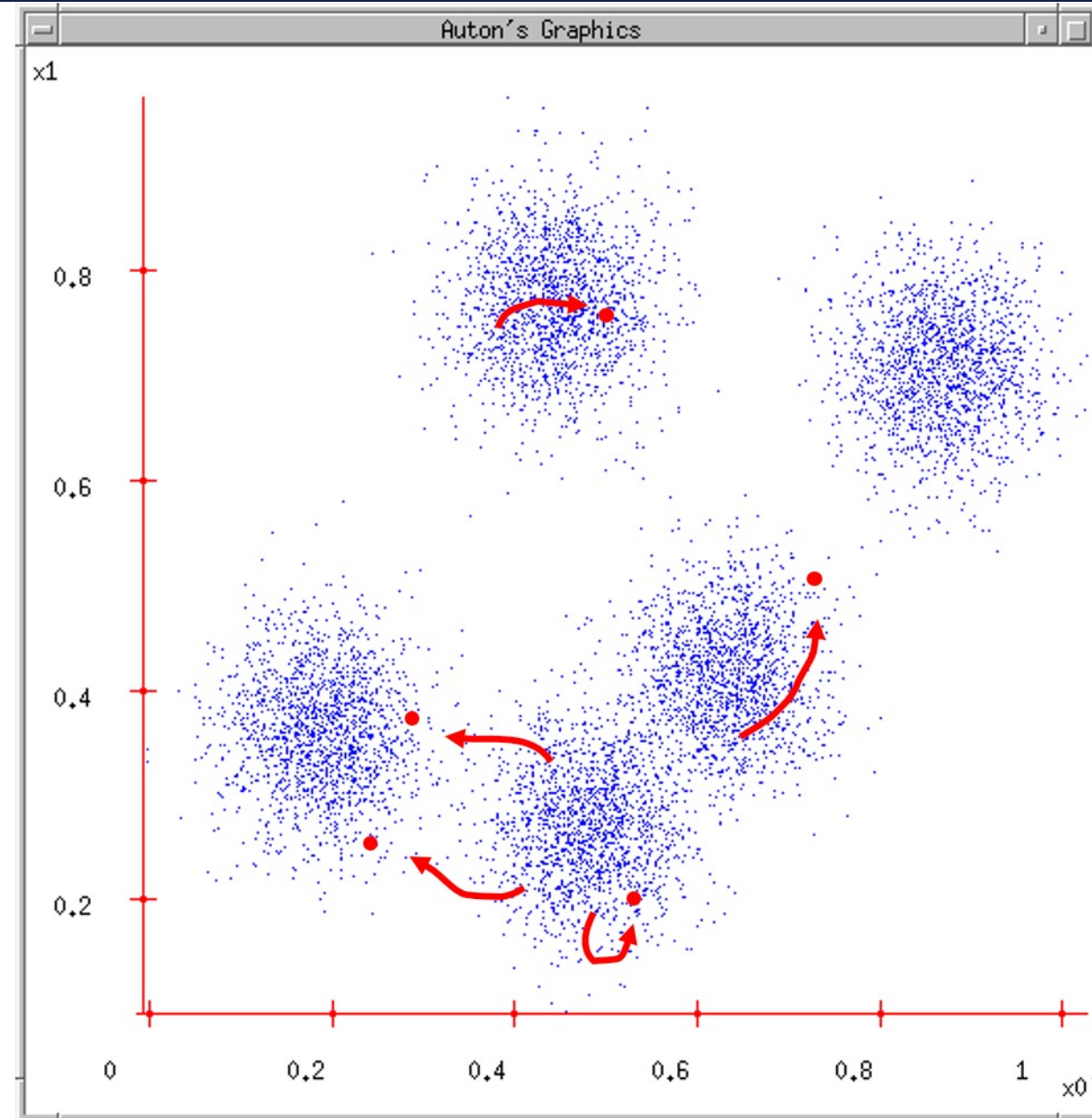4. Each Center finds the centroid of the points it owns

5. …and jumps there

6. …Repeat until terminated!

K-means Clustering

Clustering:

- *Clustering Algorithms*

  - *K-means*

1: Select K points as the initial centroids.

2: **repeat**

3:    Form K clusters by assigning all points to the closest centroid.

4:    Recompute the centroid of each cluster.

5: **until** The centroids don't change

Clustering:

- *Clustering Algorithms*

  - *K-means*

    1. Partitional clustering approach
    2. Each cluster is associated with a centroid (center point)
    3. Each point is assigned to the cluster with the closest centroid
    4. Number of clusters, K, must be specified
    5. The basic algorithm is very simple

Clustering:

- *Clustering Algorithms*

    - *K-means*

        - Strengths
            - Simple iterative method
            - User provides "K"

        - Weaknesses
            - Often too simple → bad results
            - Difficult to guess the correct "K"

K-means Clustering

Details:

❑ *Initial centroids are often chosen randomly.*
- *Clusters produced vary from one run to another.*

❑ *The centroid is (typically) the mean of the points in the cluster.*

❑ *'Closeness' is measured by Euclidean distance, cosine similarity, correlation, etc.*

❑ Iterate:
- Calculate distance from objects to cluster centroids.
- Assign objects to closest cluster
- Recalculate new centroids

❑ Stop based on convergence criteria
- No change in clusters
- Max iterations

❑ *Complexity is O(n * K * I * d)*
  n = number of points, K = number of clusters,
  I = number of iterations, d = number of attributes

UNC
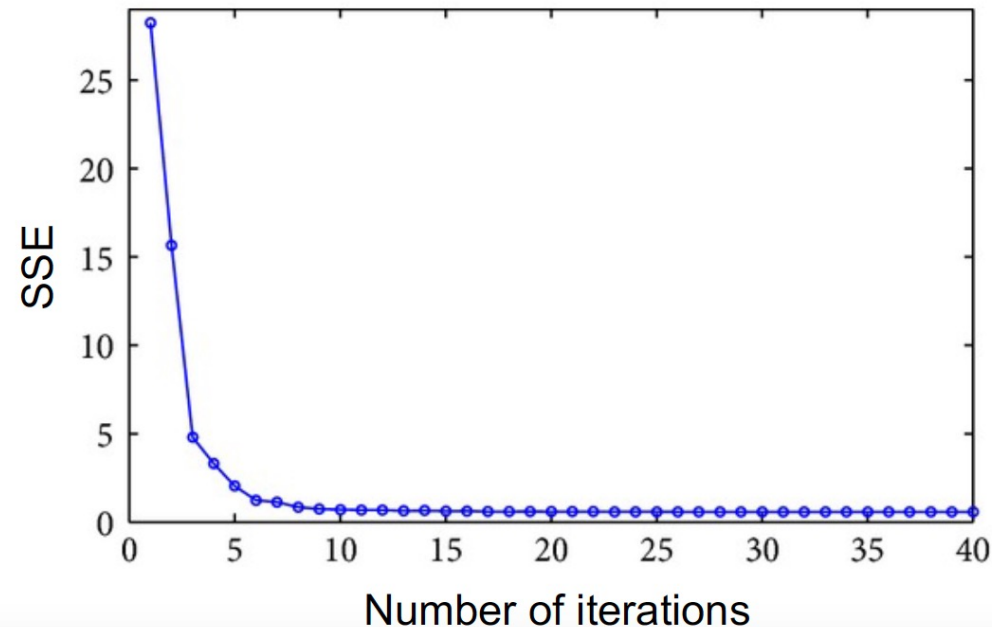GREENSBORO

K-means Clustering

Objective function in K-means

Sum of Squared Errors (SSE)
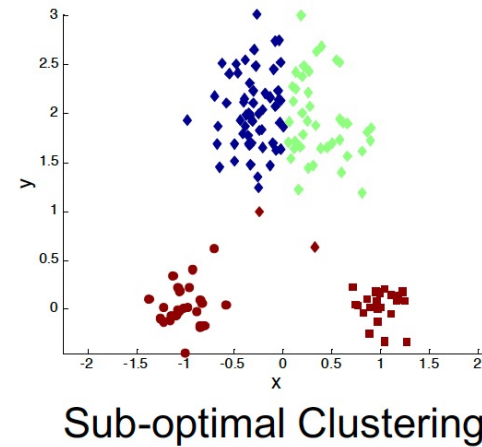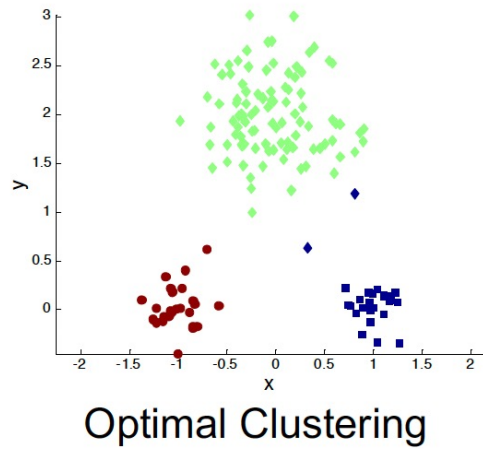
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$
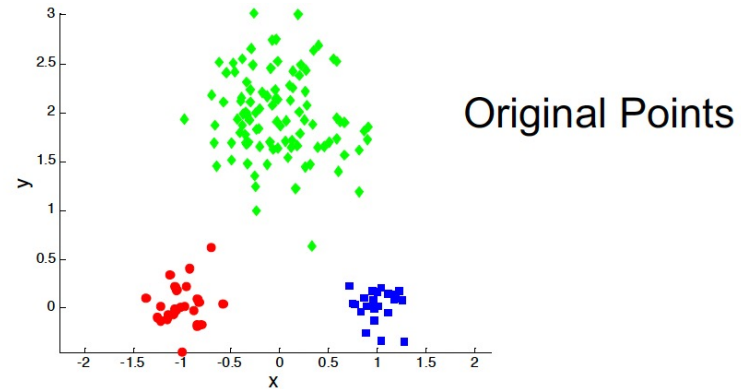
x is a data point in cluster C_i and m_i is the centroid of cluster C_i

Convergence Curve

K-means Clustering

Determine numbers of clusters

Original Points

Optimal Clustering

Sub-optimal Clustering

Choose the one with the smallest SSE

K-means Clustering

Determine numbers of clusters

Optimal number of clusters
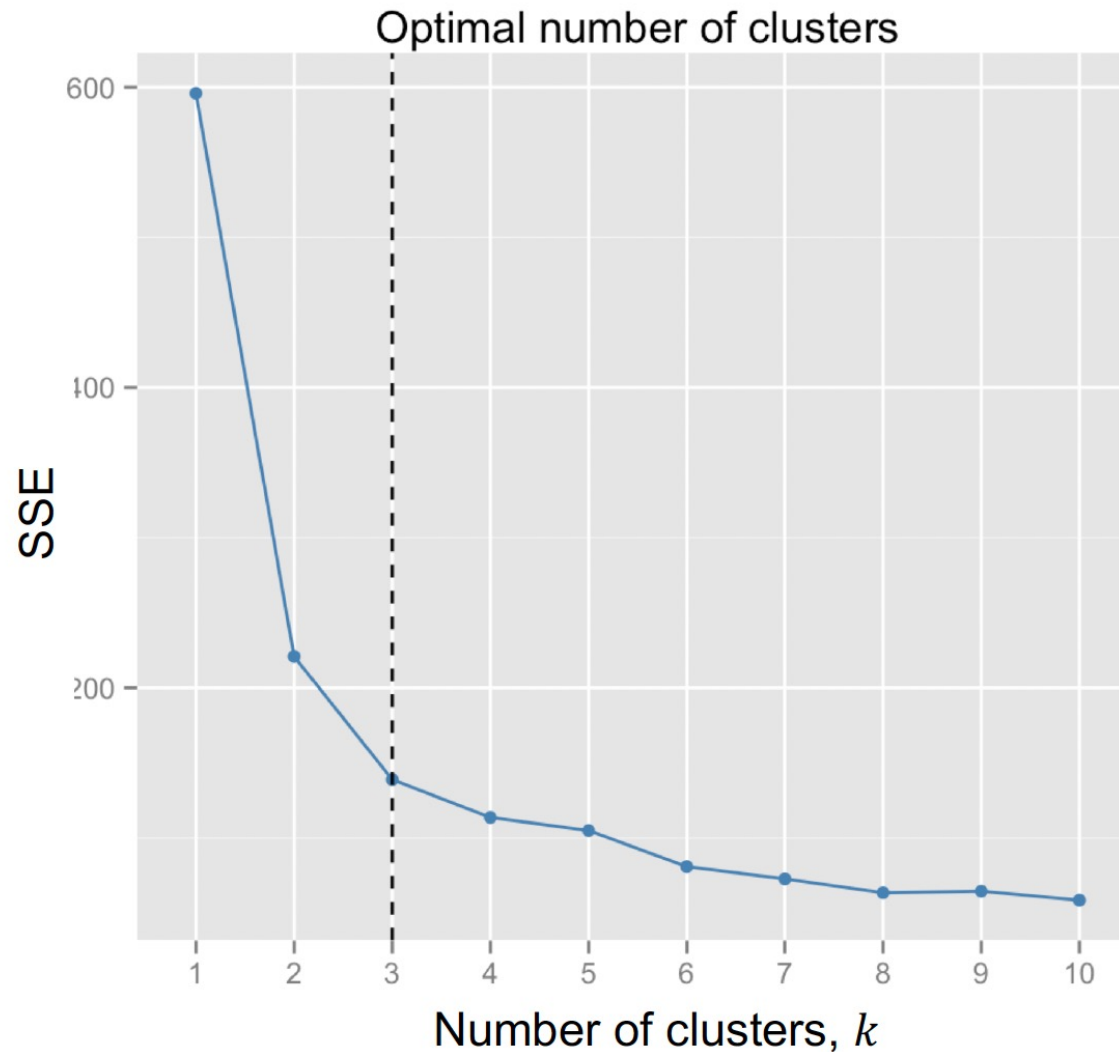
K-means Clustering

How can you do this efficiently?

- Idea 1: Be careful about where you start
- Idea 2: Do many runs of k-means, each from a different random start configuration
- Many other ideas floating around.

How can you do this efficiently?

- Idea 1: Be careful about where you start
- Idea 2: Do ~~~~ runs of k-means, each from a diff~~~~
- Ma~~~~

Neat trick:
Place first center on top of randomly chosen datapoint.
Place second center on datapoint that's as far away as possible from first center
    :
Place j'th center on datapoint that's as far away as possible from the closest of Centers 1 through j-1
    :

UNC
GREENSBORO

Codes available on github:

https://github.com/q-tong/CS405-605-Data-Science/blob/main/lecture/4.%20Clustering-KMeans.ipynb