STEVENS INSTITUTE of TECHNOLOGY
School of Business

Multivariate Data Analytics

Clustering Analysis

Prof. Feng Mai
School of Business

For academic use only.

Clustering

- **Objective:** Dividing records (rows) in a multivariate dataset into “natural” clusters (groups), where the records in each group are similar to one another.
- **Applications**
  - Cluster customers into different segments
  - Cluster products according their attributes
  - Cluster businesses according to their location
- **Methods:** K-means, Model-based Clustering (Gaussian Mixture Model), Hierarchical Clustering, DBSCAN...
- **Distance Measure**
  - Euclidean Distance (most common): square-root of the sum of the squared differences between each variable.

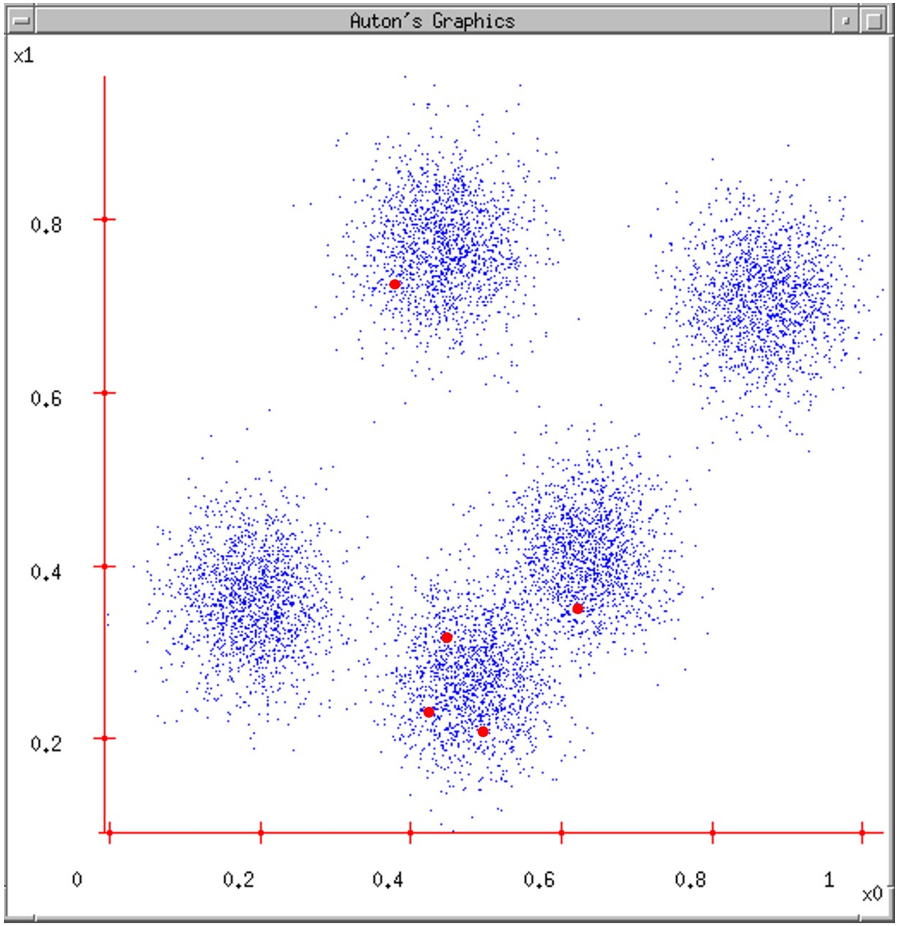
$$d(\mathbf{X_i}, \mathbf{X_j}) = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}$$

  - Other measures may be more appropriate for a specific dataset

K-Means

K-means

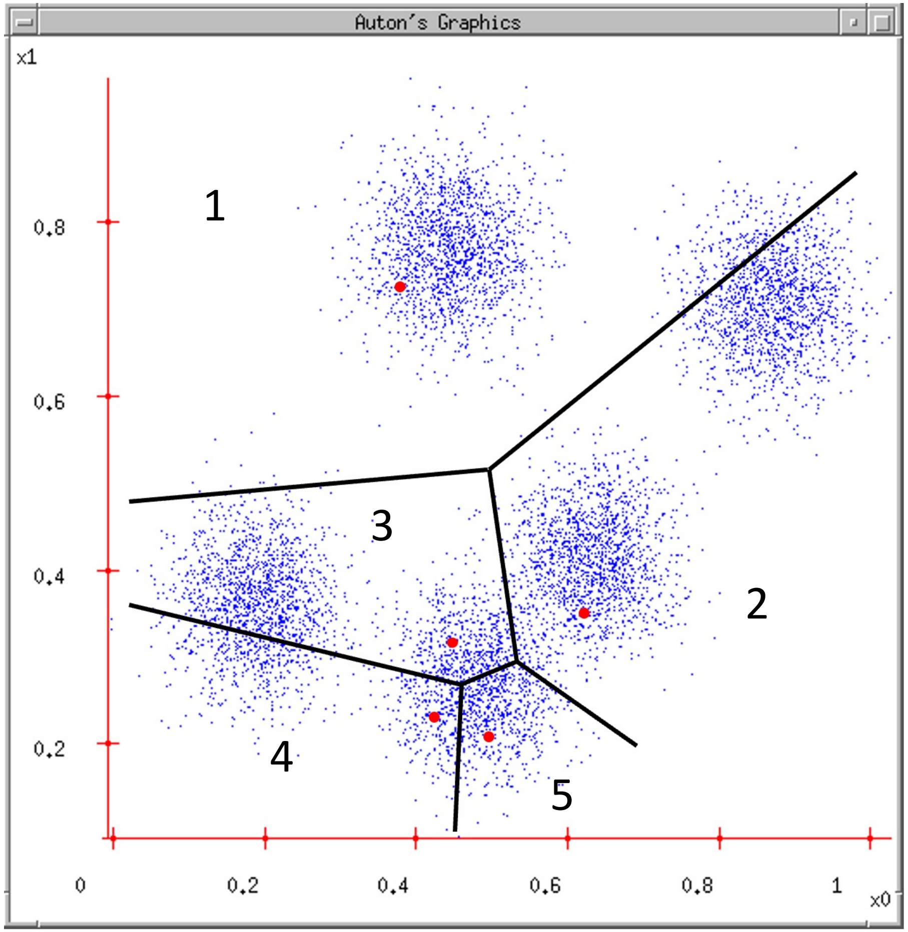
1. Ask user how many clusters they'd like. (k = 5)
2. Randomly guess k cluster centers



Example by Andrew W. Moore

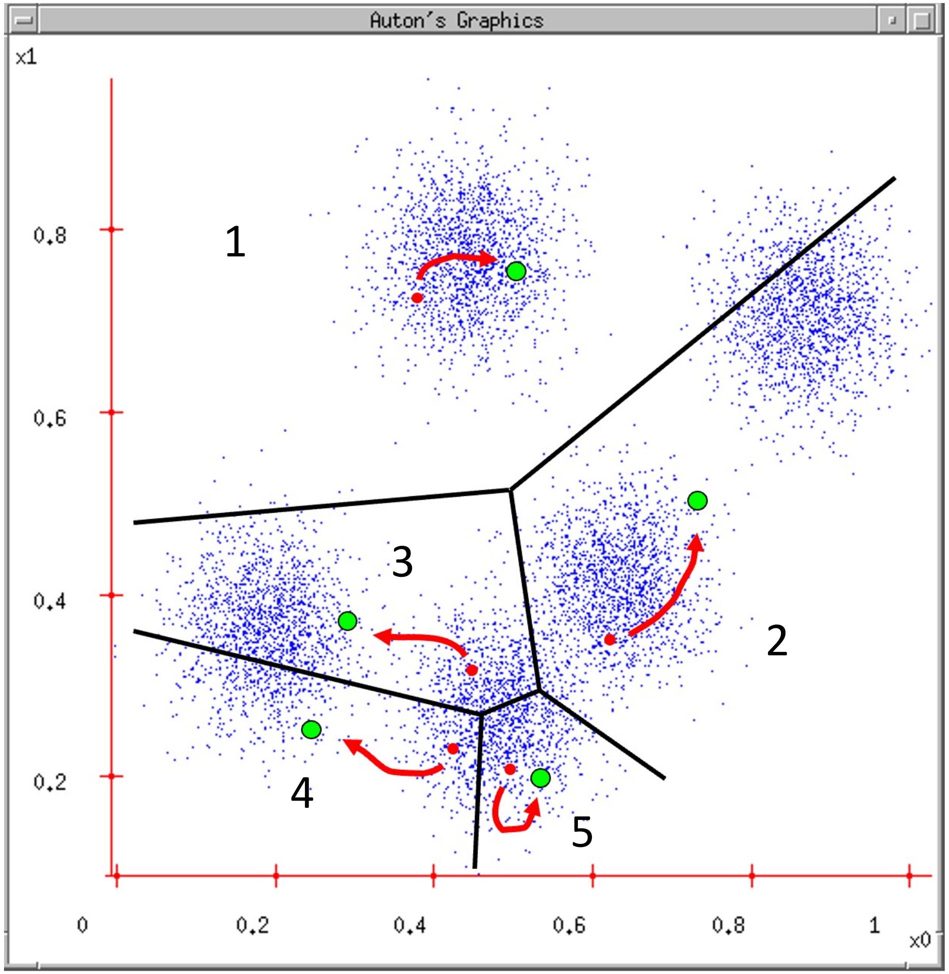
K-means

1. Ask user how many clusters they'd like. (k = 5)
2. Randomly guess k cluster Centers
3. Assign blue datapoint to closest Center. (each Center "owns" a set of datapoints)



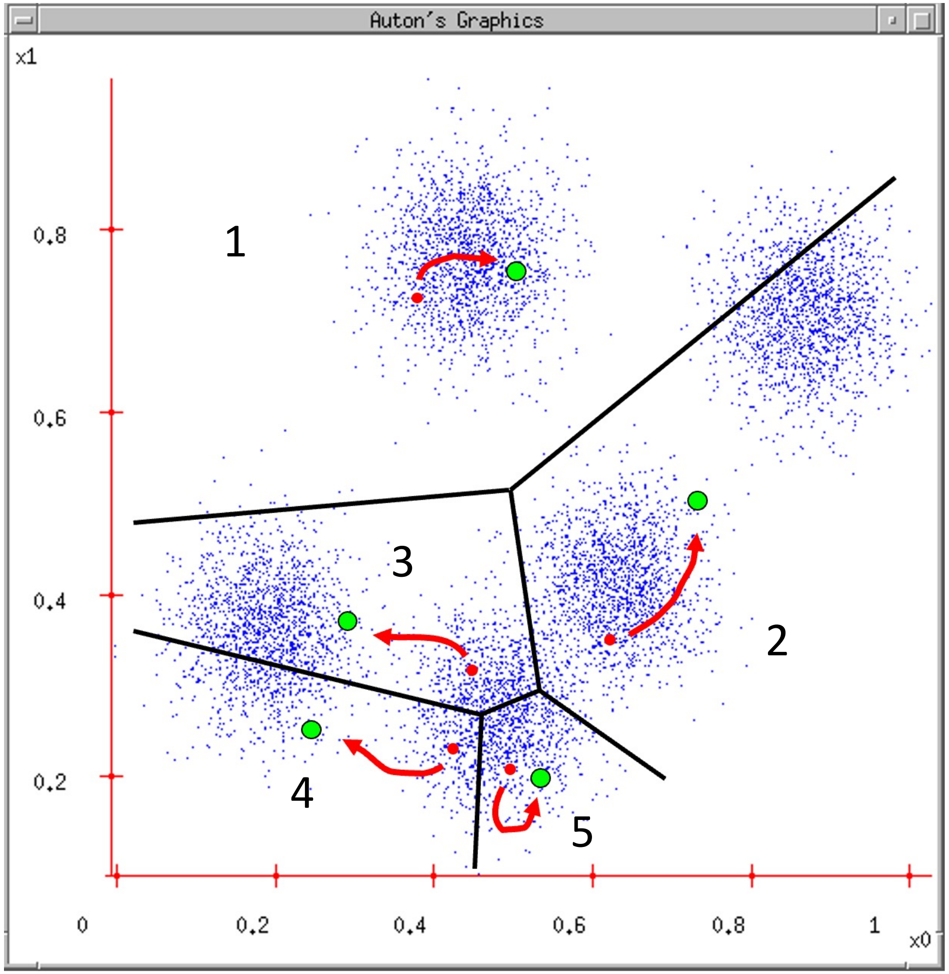
K-means

1. Ask user how many clusters they'd like. (k = 5)
2. Randomly guess k cluster Centers
3. Assign blue datapoint to closest Center. (each Center "owns" a set of datapoints)
4. Each **Old Center** moves to the **New Center** of the points it owns

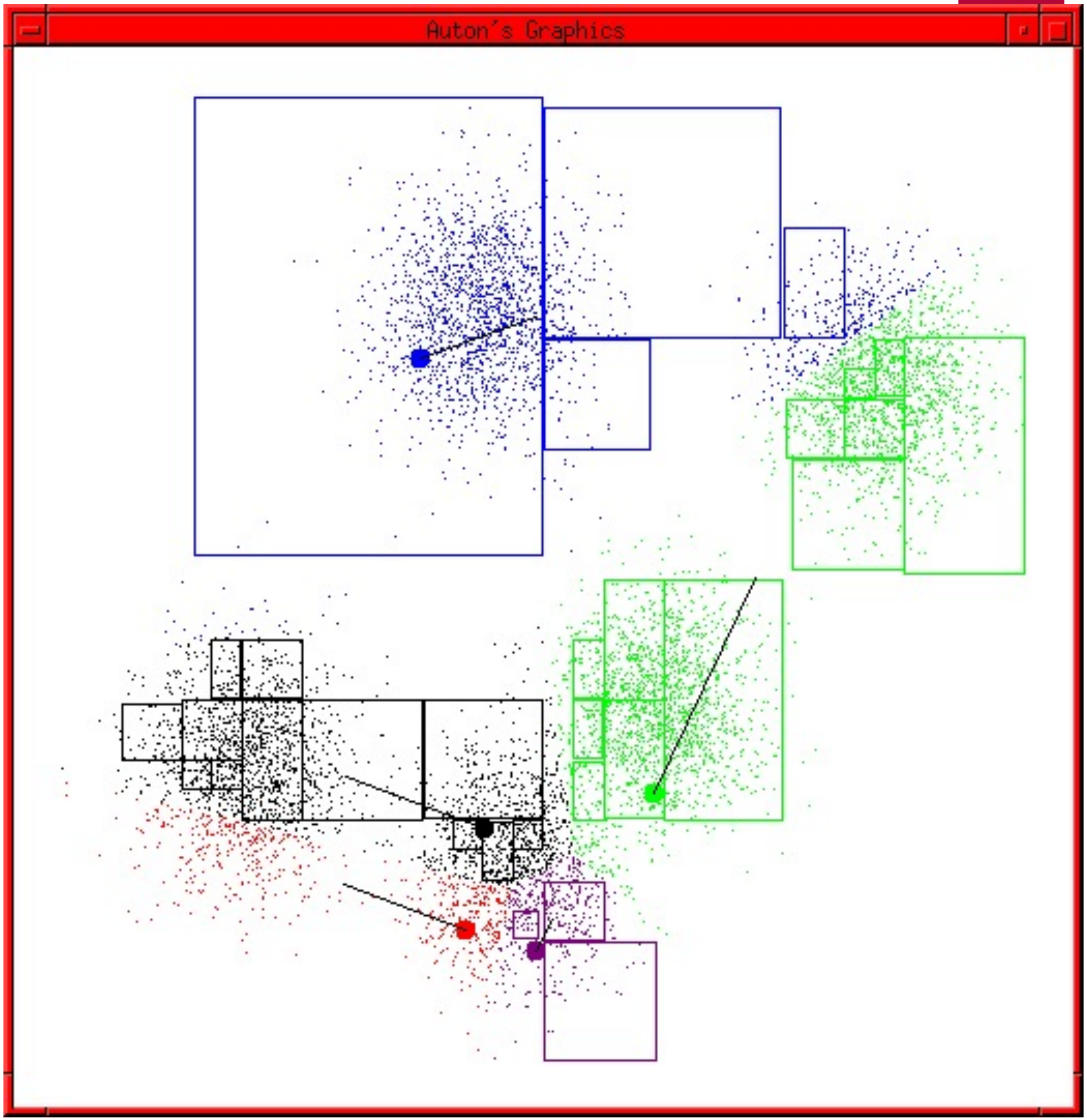


K-means

1. Ask user how many clusters they'd like. (k = 5)
2. Randomly guess k cluster Centers
3. Assign blue datapoint to closest Center. (each Center "owns" a set of datapoints)
4. Each Old Center moves to the New Center of the points it owns
5. Jump to Step 3. Repeat until terminated.

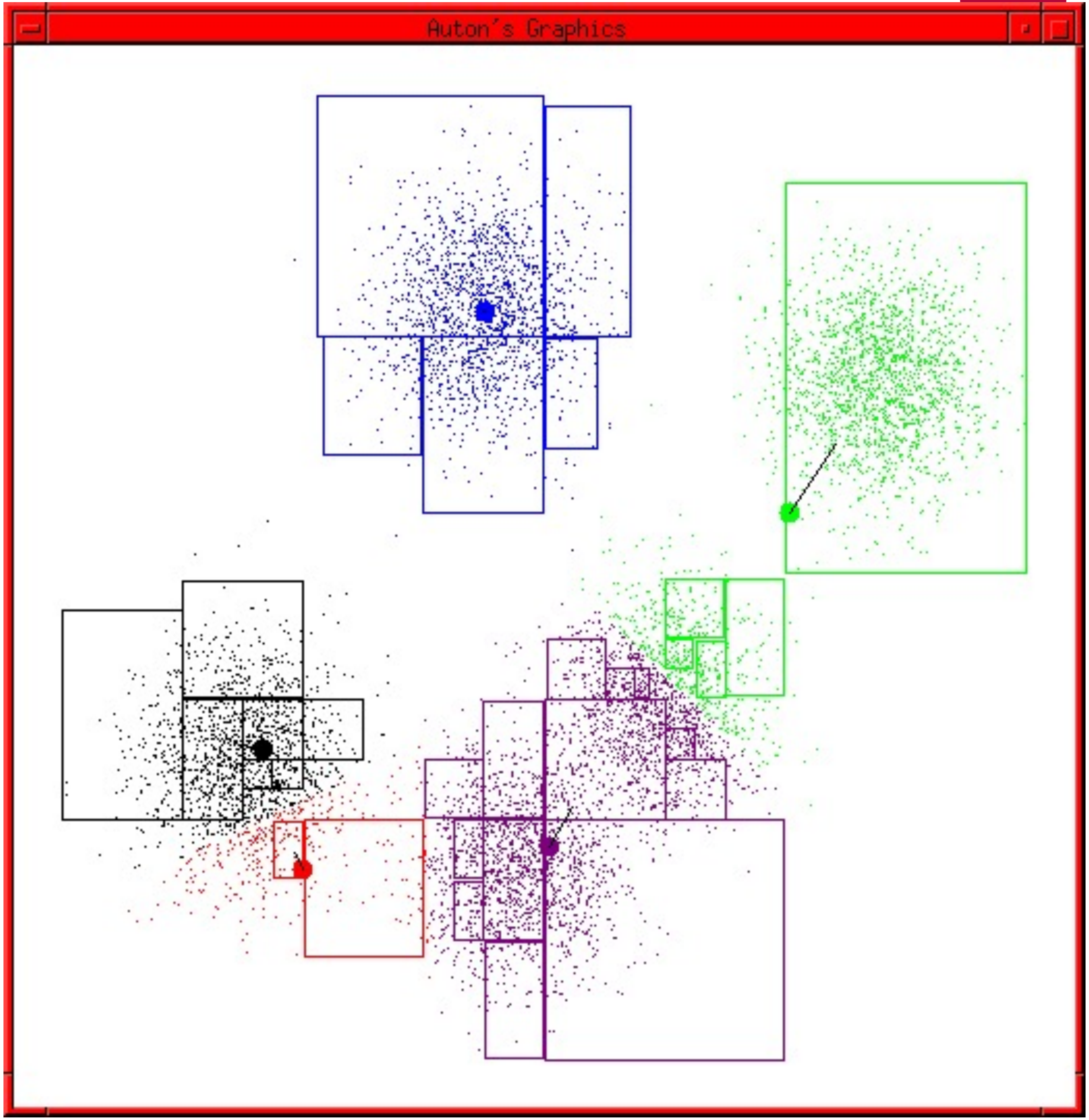


K-means starts

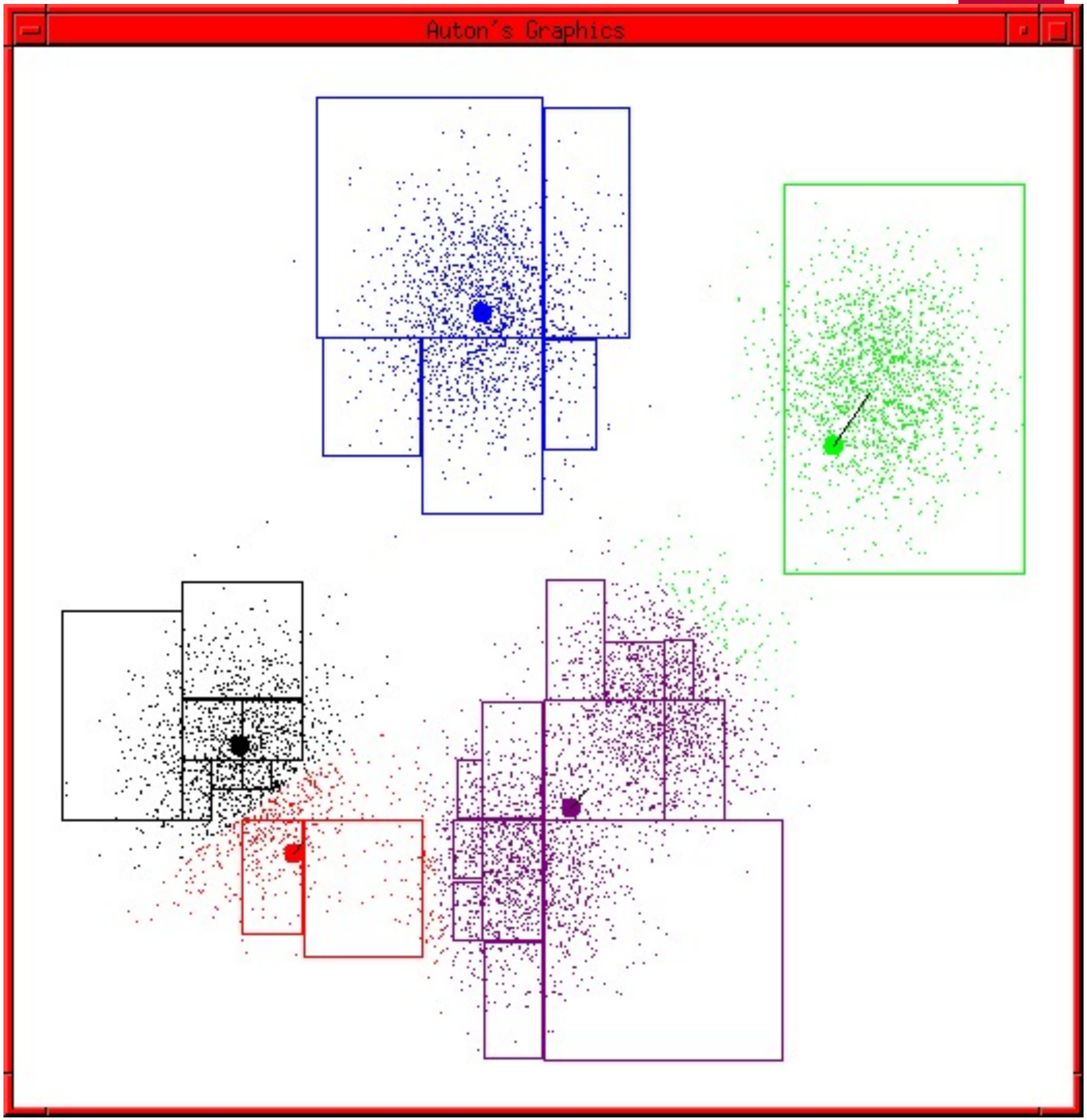


Source: www.autonlab.org

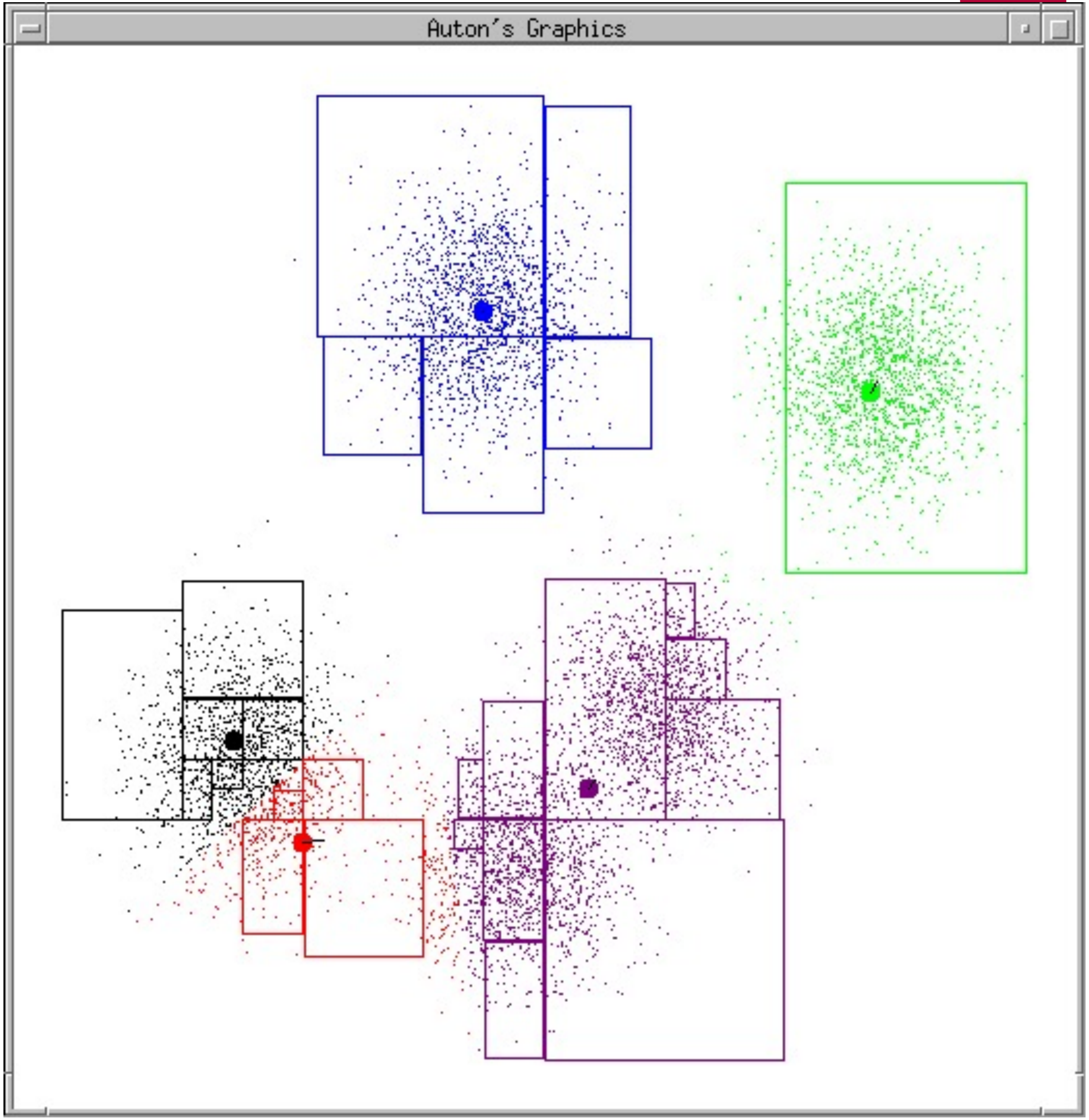
K-means continues…



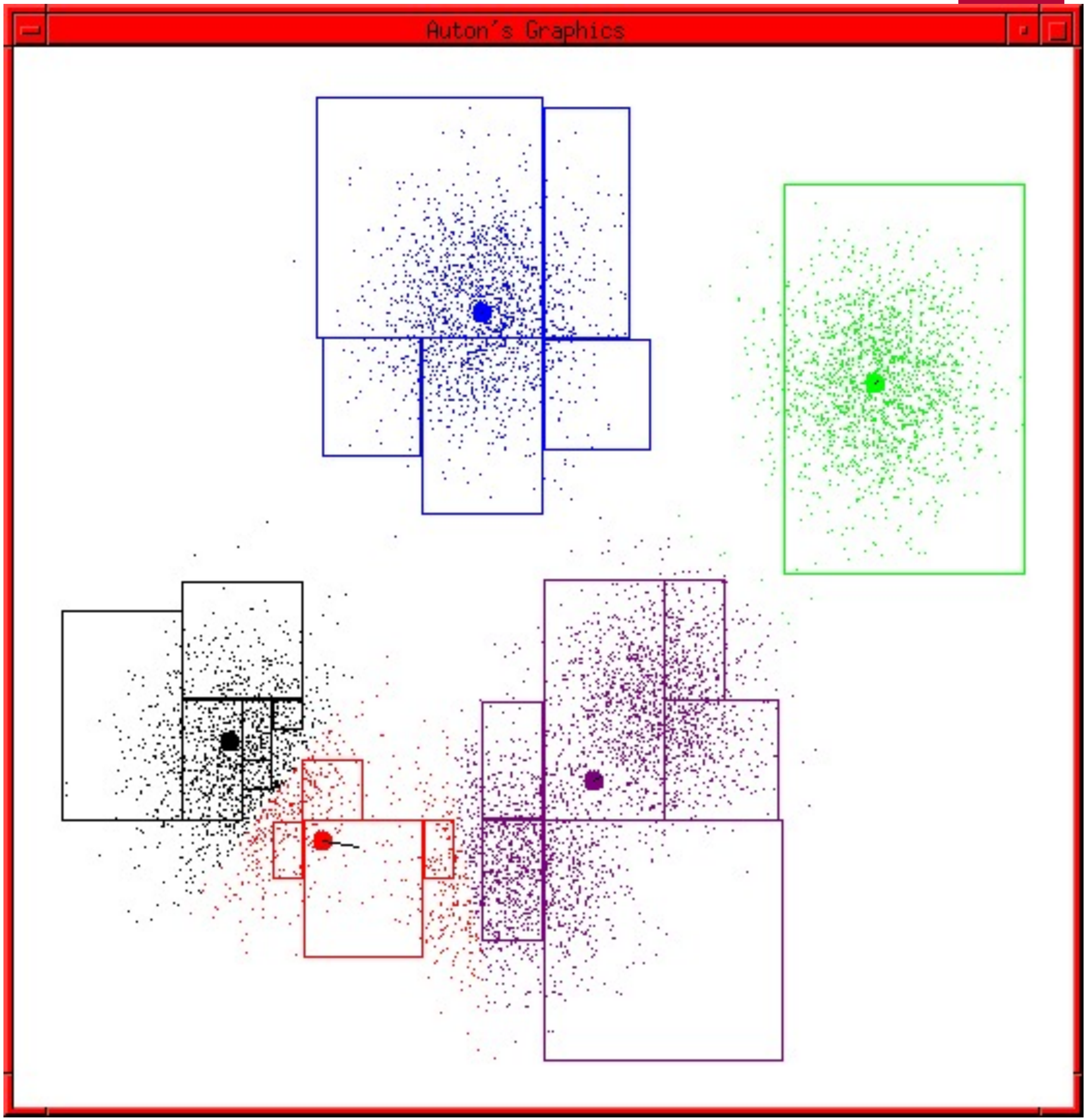
K-means continues…



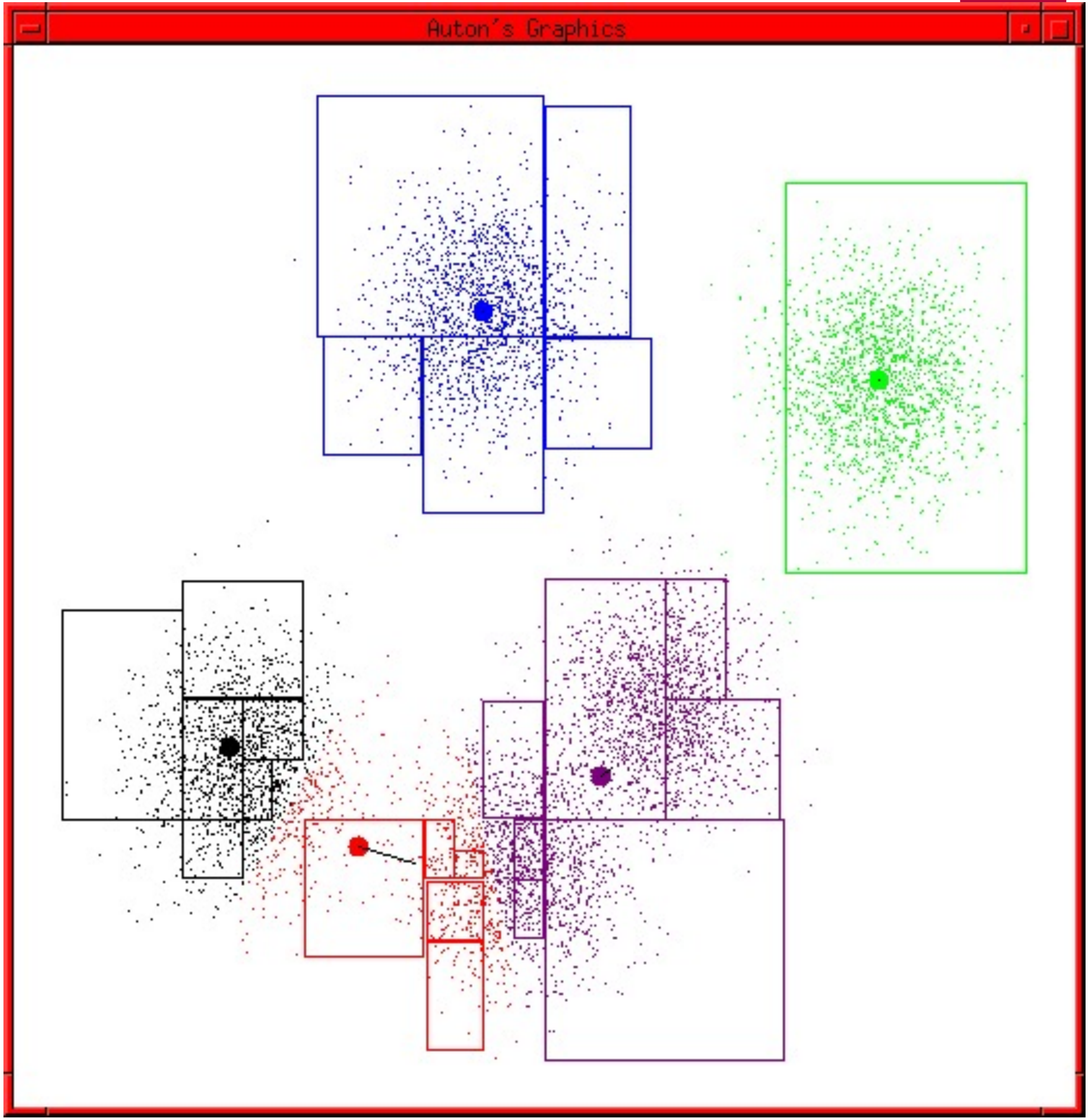
K-means continues…



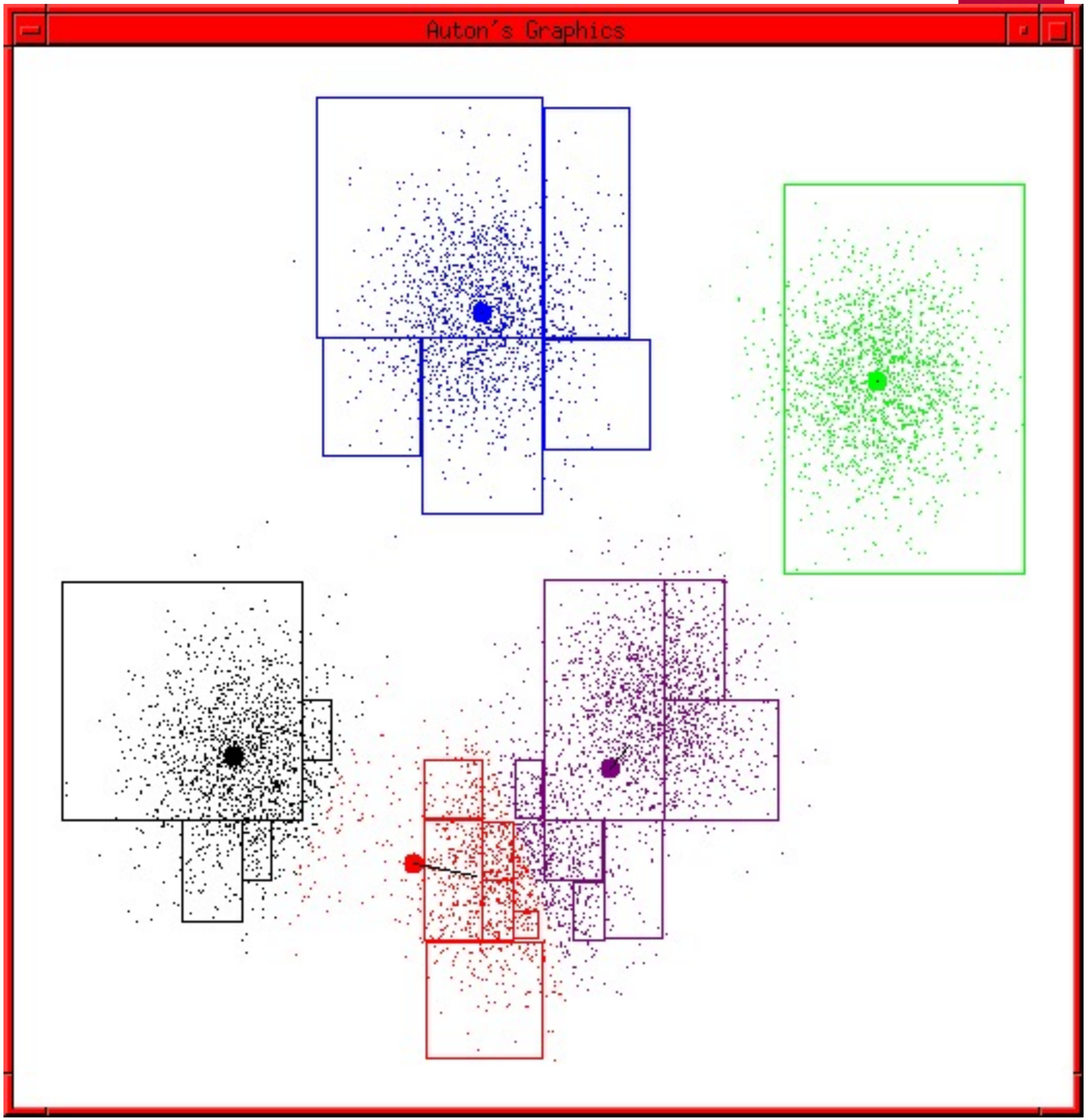
K-means continues…



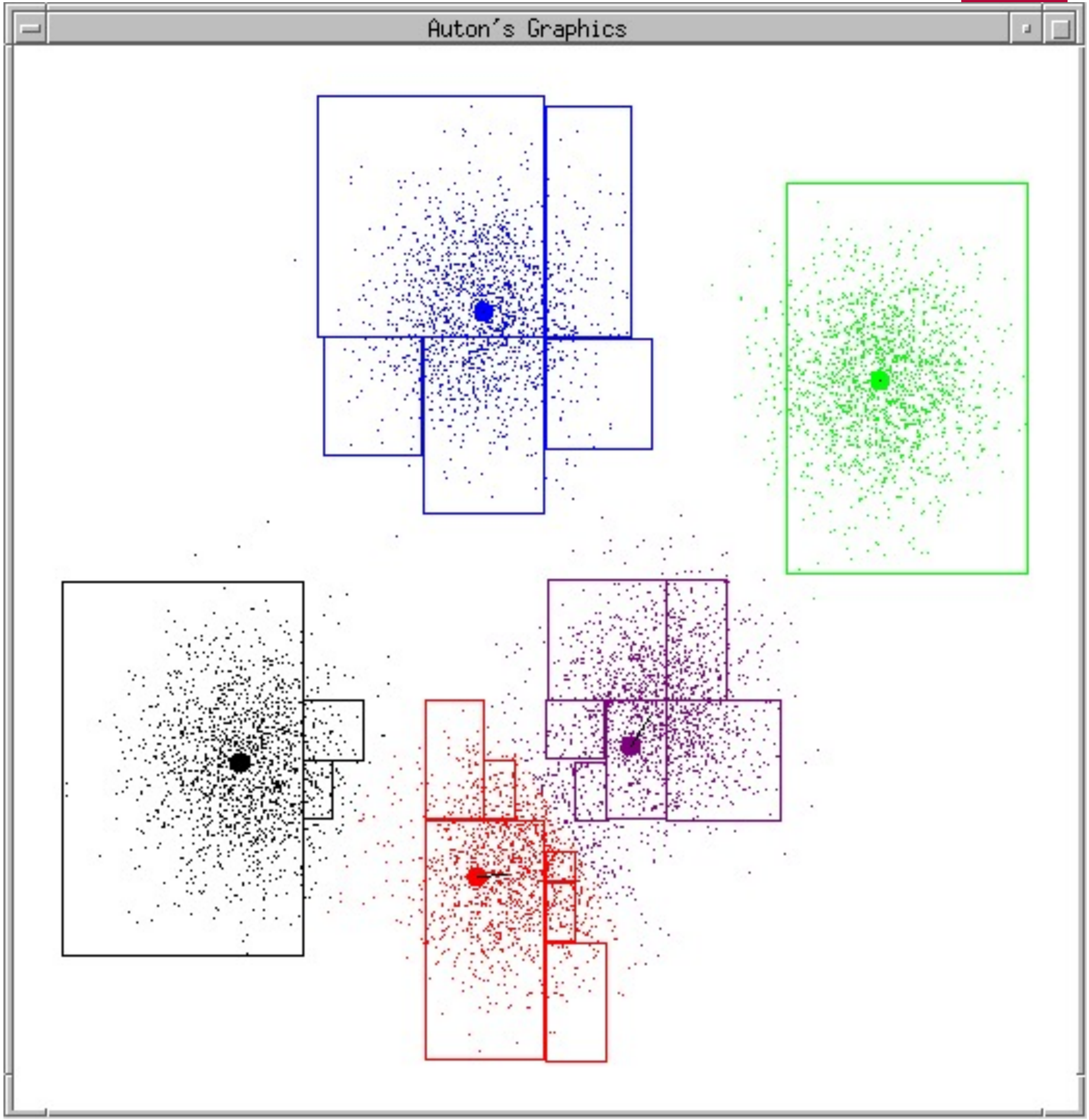
K-means continues…



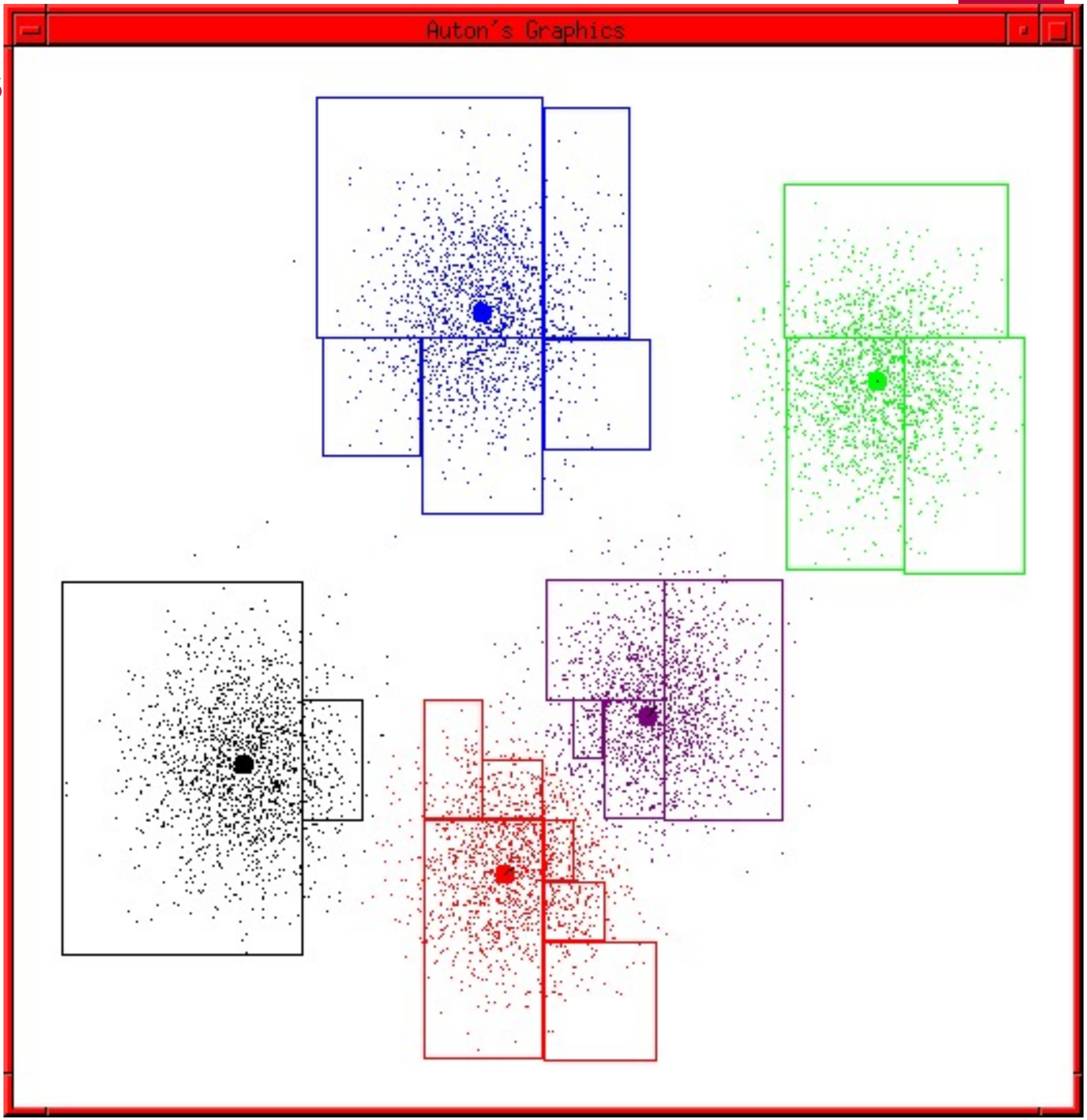
K-means continues…



K-means continues…



K-means terminates



K-means Questions

- What is it trying to optimize?
- Are we sure it will terminate?
- Are we sure it will find an optimal clustering?
- How should we start it?
- How could we automatically choose the number of centers?

K-means objective

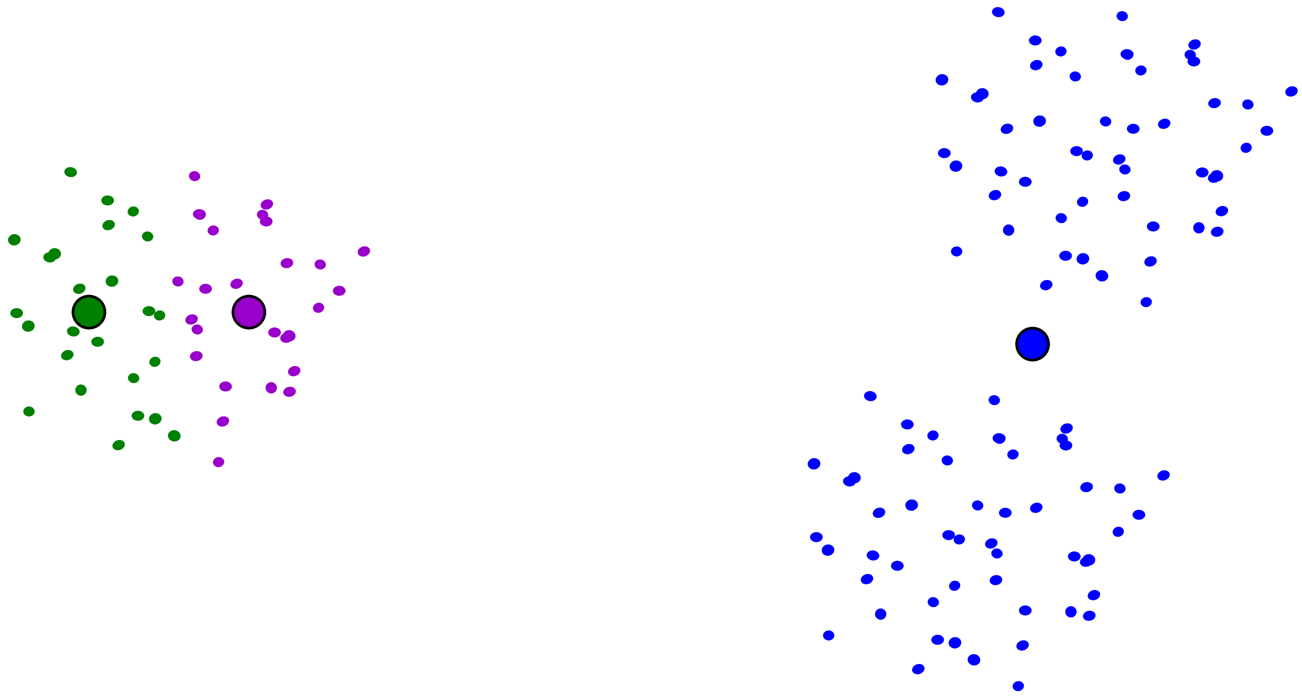
Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance). Formally, the objective is to find:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2 = \underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

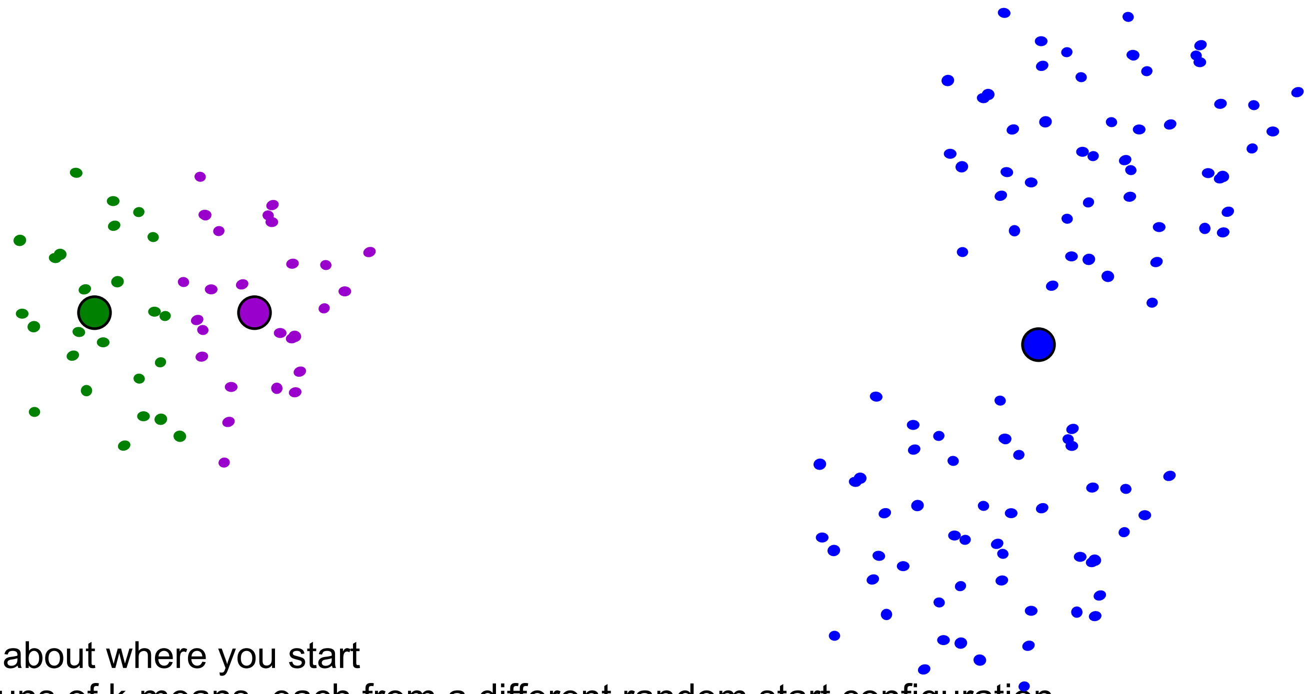
Will we find the optimal configuration?

Not necessarily.

Will we find the optimal configuration?



Trying to find good optima



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. For example:

- 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

Choosing the number of clusters (k)

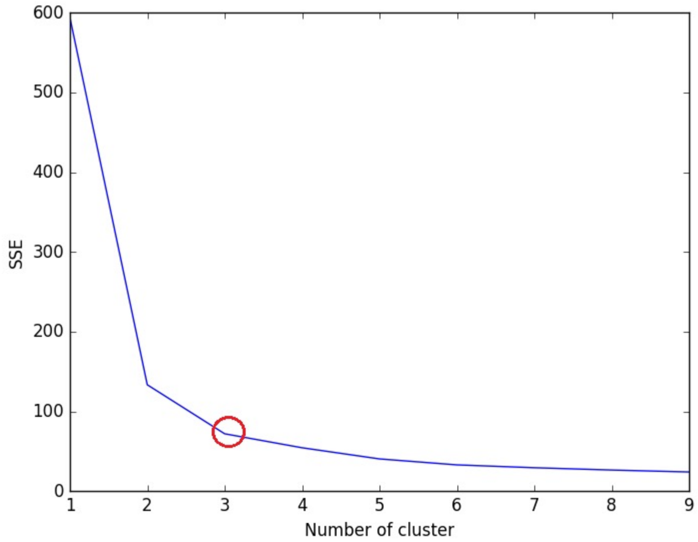
- No simple answer
- Subjective evaluation: Are clusters interpretable?
- Elbow method:

Y-axis: **sum of squared errors (SSE)** inside each cluster (the squared difference between points and their cluster center). Alternatively, use **Variance Explained%.**

X-axis: number of clusters

Determine the elbow in the graph

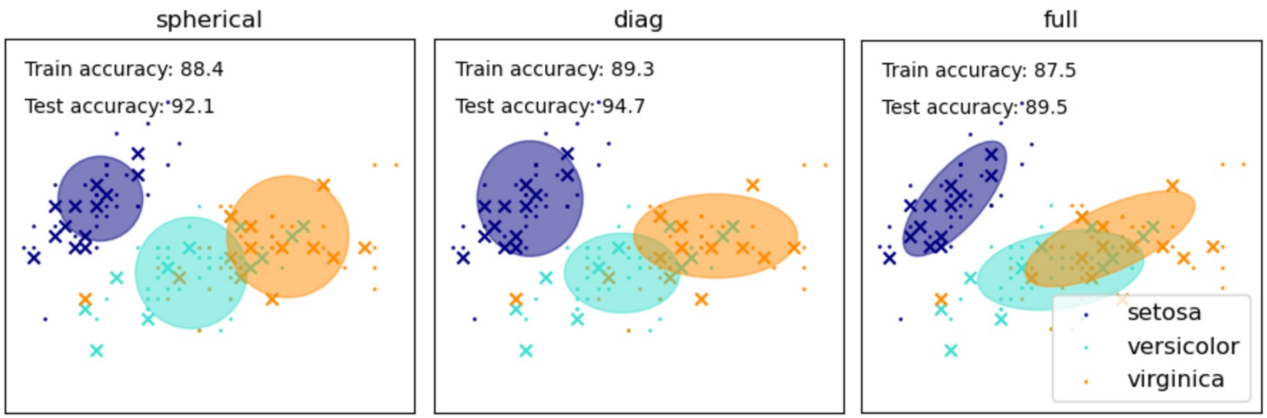
- Other quality measures:
  - e.g., Silhouette coefficient, Dunn index...



Other Clustering Methods

Model-Based Clustering

- Grounded in Multivariate Normal (Gaussian) Distribution
  - The data is generated from Mixtures of Normal
  - Constraining the covariance matrix to restrict the shape of clusters
- Can choose the number of clusters using criterion such as BIC
- Provides “soft” or probabilistic cluster assignment
- Uses the Expectation-Maximization (EM) algorithm to maximize the Likelihood function (see readings)
- K-means can be considered as a special case

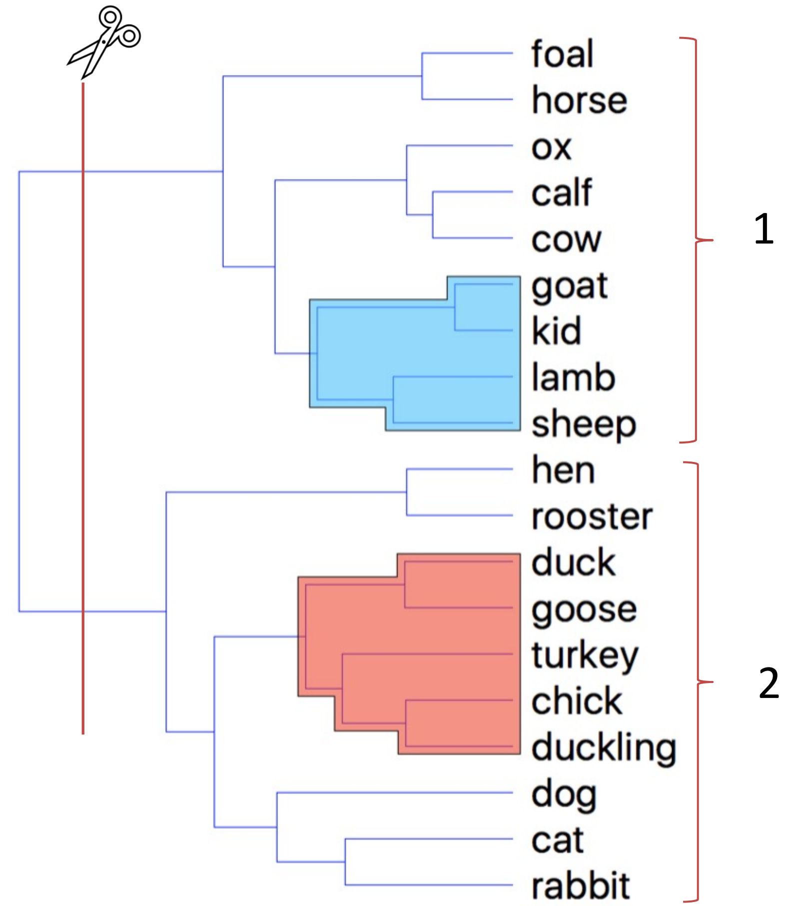


Source: https://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_covariances.html

Hierarchical Clustering

- Merges or splits records in a greedy manner
  - Agglomerative**:** each observation starts in its own cluster, and pairs of clusters that are most similar to each other are merged
  - Divisive: all observations start in one cluster and clusters are spitted based on similarity
- Allows flexible distance (similarity) measure; not limited to Euclidean Distance
- Produces a **dendrogram** (tree) for the records
- We can choose the number of clusters by cutting the dendrogram at any level

2-cluster solution

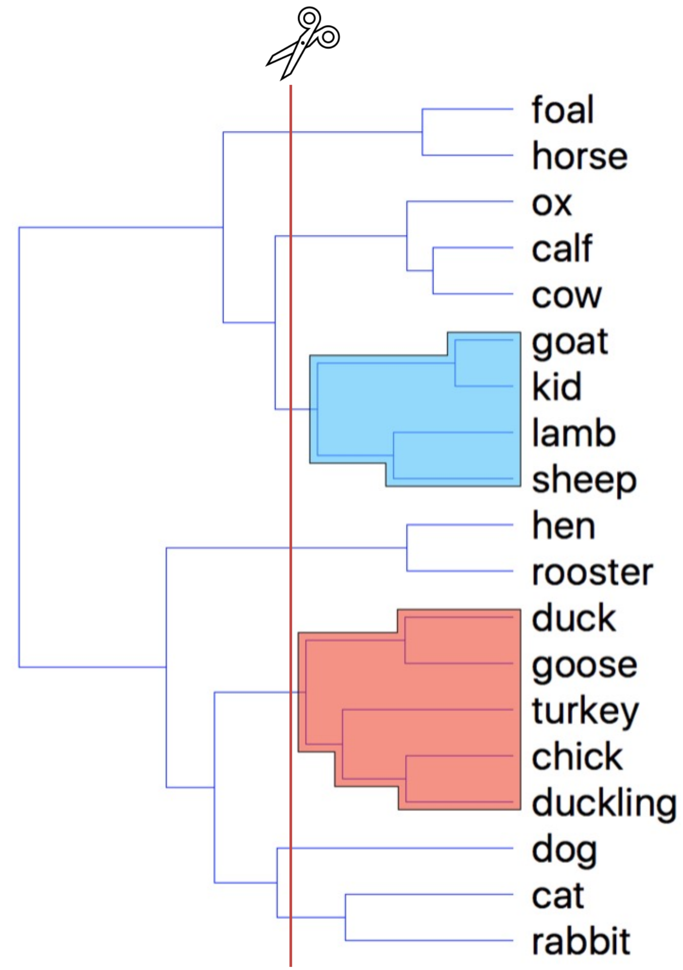


Source: https://en.wikipedia.org/wiki/File:Orange-data-mining-hierarchical-clustering.png

Hierarchical Clustering

- Merges or splits records in a greedy manner
  - Agglomerative**:** each observation starts in its own cluster, and pairs of clusters that are most similar to each other are merged
  - Divisive: all observations start in one cluster and clusters are spitted based on similarity
- Allows flexible distance (similarity) measure; not limited to Euclidean Distance
- Produces a **dendrogram** (tree) for the records
- We can choose the number of clusters by cutting the dendrogram at any level

7-cluster solution



Source: https://en.wikipedia.org/wiki/File:Orange-data-mining-hierarchical-clustering.png

DBSCAN (Density-based spatial clustering of applications with noise)

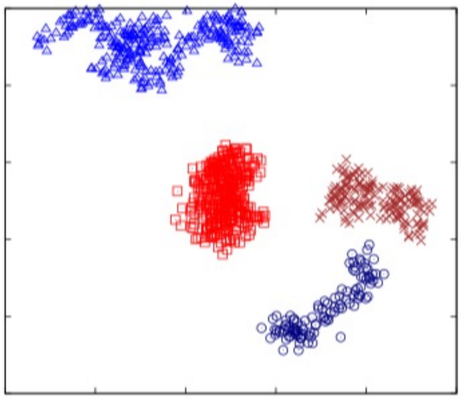
- Groups together points with many nearby neighbors
  - Neighbor is defined using a distance threshold ϵ

Advantages

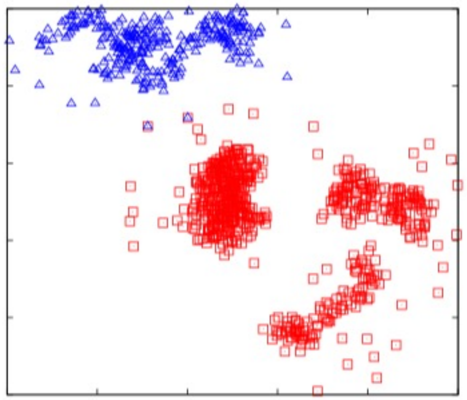
- Finds arbitrarily-shaped clusters
- No need to choose the number of clusters
- Handles outliers (finds noise points that do not belong a cluster)
- Fast for certain databases

Disadvantage

- Can be sensitive to the distance threshold ϵ



(d) $\rho = 0.1, \epsilon = 5000$



(h) $\rho = 0.1, \epsilon = 11300$

Source: Gan, J., & Tao, Y. (2015). DBSCAN revisited: Mis-claim, un-fixability, and approximation. *ACM SIGMOD*

Thank you!

Prof. Feng Mai
School of Business