Fall 2018: Introduction to Data Science

GIRI NARASIMHAN, SCIS, FIU

Clustering

Clustering dogs using height & weight

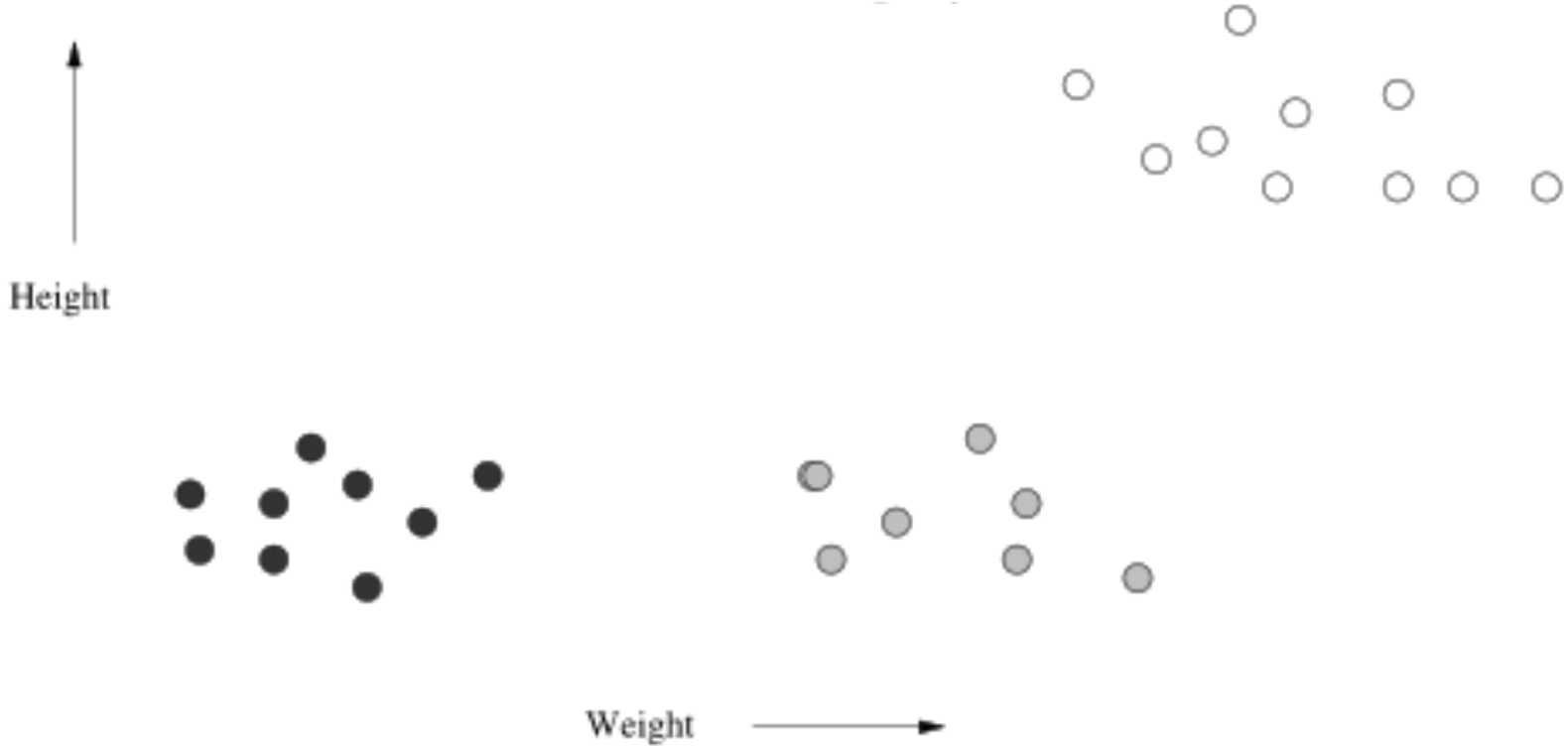


Figure 7.1: Heights and weights of dogs taken from three varieties

Clustering dogs using height & weight

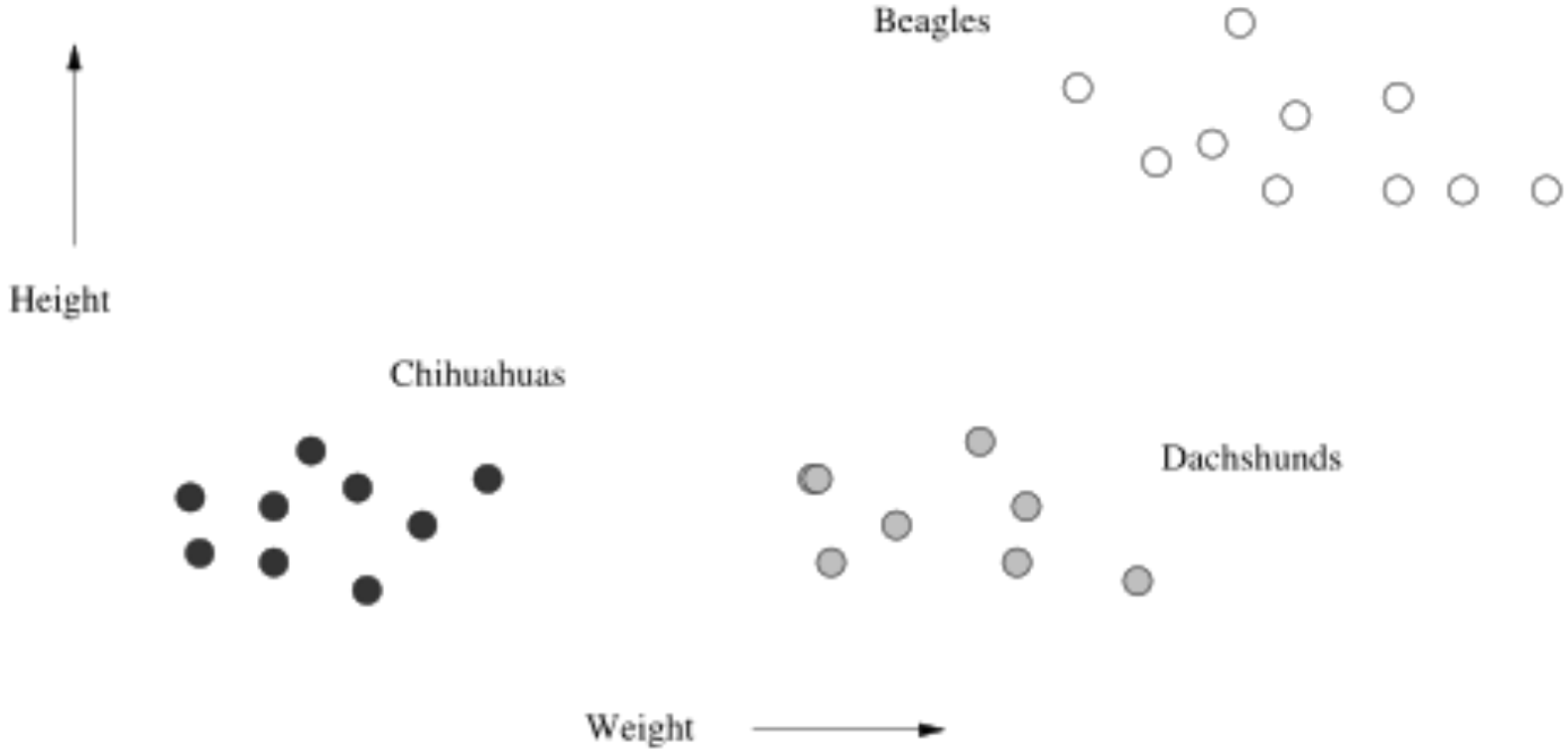


Figure 7.1: Heights and weights of dogs taken from three varieties

Clustering

- Clustering is the process of making clusters, which put **similar** things together into same cluster …
- And put **dissimilar** things into different clusters
- Need a similarity function
- Need a ~~similarity~~ **distance** function
  - Convenient to map items to points in space

Distance Functions

- Jaccard Distance
- Hamming Distance
- Euclidean Distance
- Cosine Distance
- Edit Distance
- …

- What is a **distance** function
  - D(x,y) >= 0
  - D(x,y) = D(y,x)
  - D(x,y) <= D(x,z) + D(z,y)

Clustering Strategies

- Hierarchical or Agglomerative
  - Bottom-up
- Partitioning methods
  - Top-down
- Density-based
- Cluster-based
- Iterative methods

Curse of Dimensionality

- N points in d-dimensional space
  - If d = 1, then average distance = 1/3
  - As d gets larger, what is the average distance? Distribution of distances?
    - # of **nearby** points for any a given point **vanishes.** So, clustering does not work well
    - # of points at max distance (~sqrt(d)) also vanishes. Real range actually very small
  - Angle ABC given 3 points approaches 90
    - Denominator grows linearly with d
    - Expected cos = 0 since equal points expected in all 4 quadrants

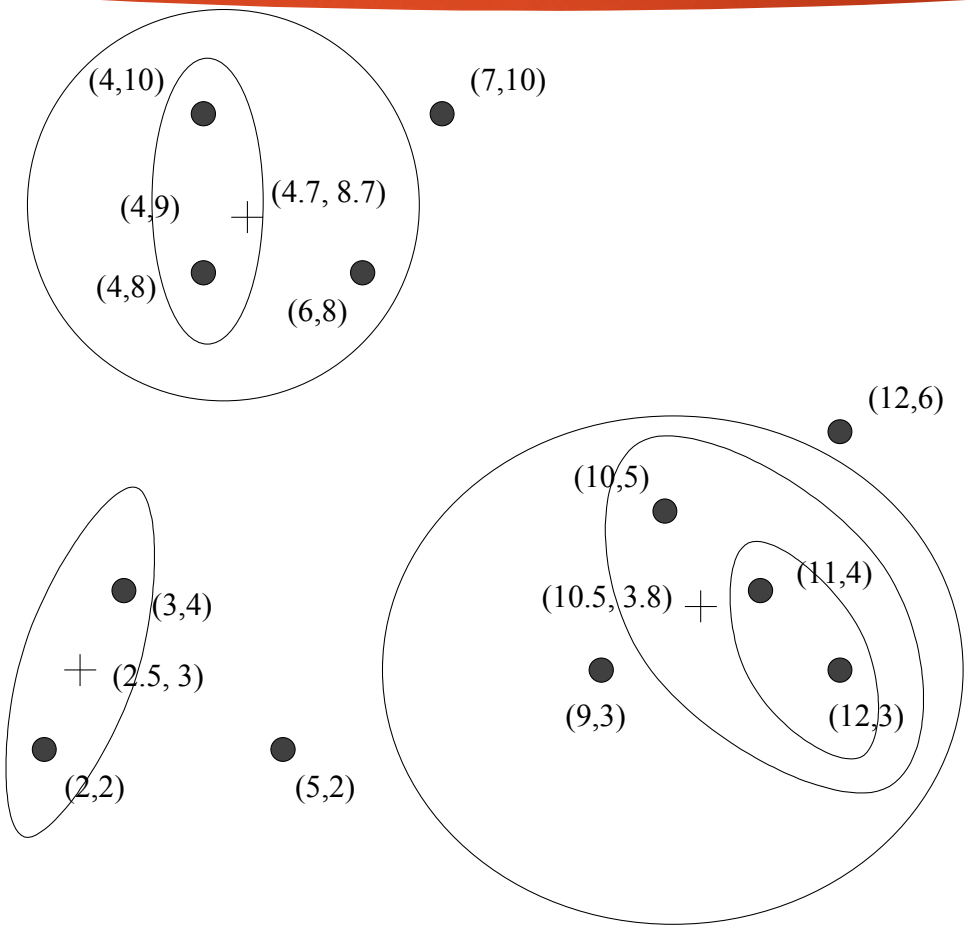
$$\frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} x_i^2}\sqrt{\sum_{i=1}^{d} y_i^2}}$$

Hierarchical Clustering

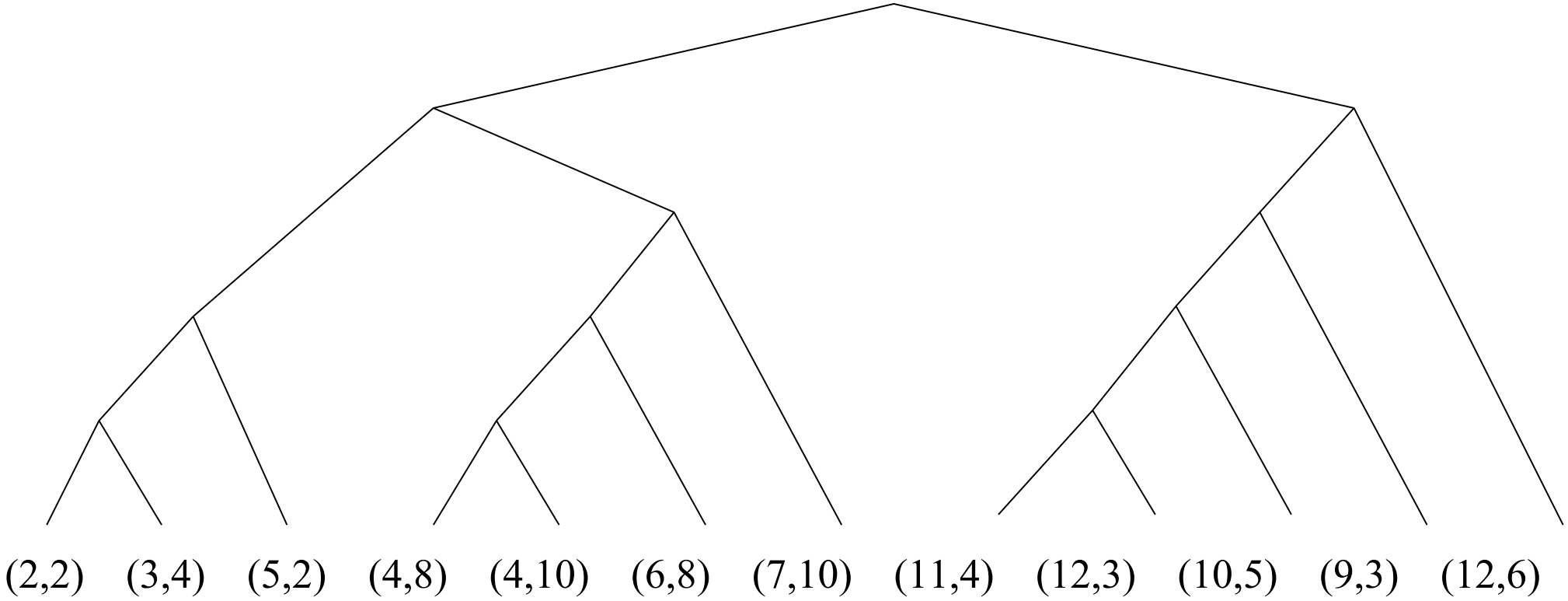
Hierarchical Clustering

- Starts with each item in different clusters
- Bottom up
- In each iteration
  - Two clusters are identified and merged into one
- Items are combined as the algorithm progresses
- **Questions**:
  - How are clusters represented
  - How to decide which ones to merge
  - What is the sopping condition
- Typical algorithm: find smallest distance between nodes of different clusters

Hierarchical Clustering



Output of Clustering: Dendrogram



Measures for a cluster

- Radius: largest distance from a centroid
- Diameter: largest distance between some pair of points in cluster
- Density: # of points per unit volume
- Volume: some power of radius or diameter

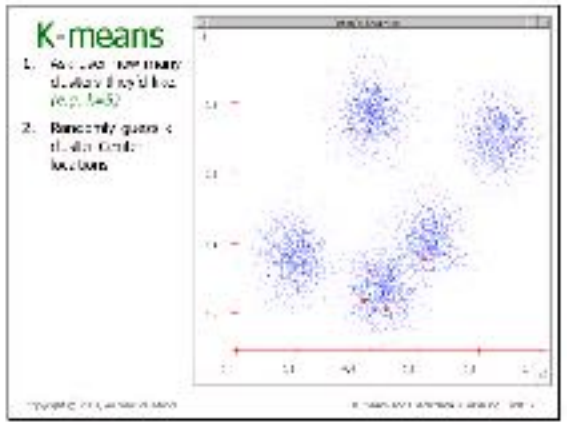
- **Good cluster**: when diameter of each cluster is much larger than its nearest cluster or nearest point outside cluster

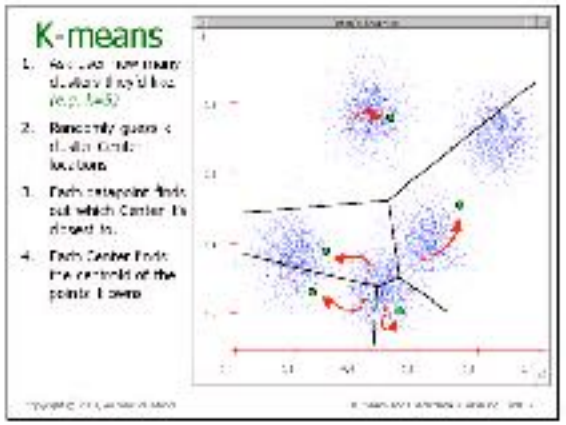
Stopping condition for clustering

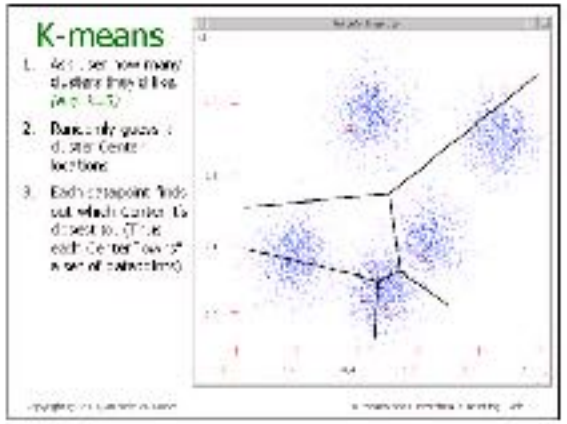
- Cluster radius or diameter crosses a threshold
- Cluster density drops below a certain threshold
- Ratio of diameter to distance to nearest cluster drops below a certain threshold

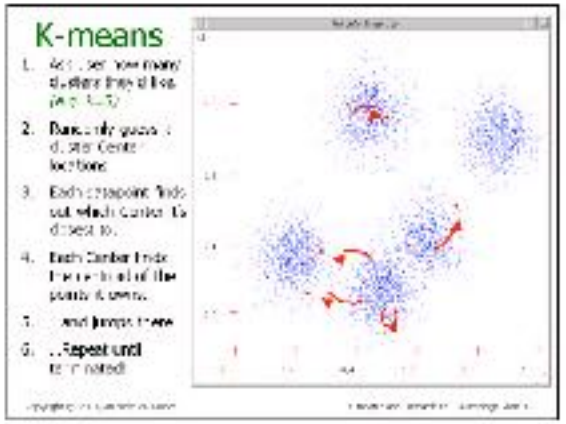
K-Means Clustering

Start





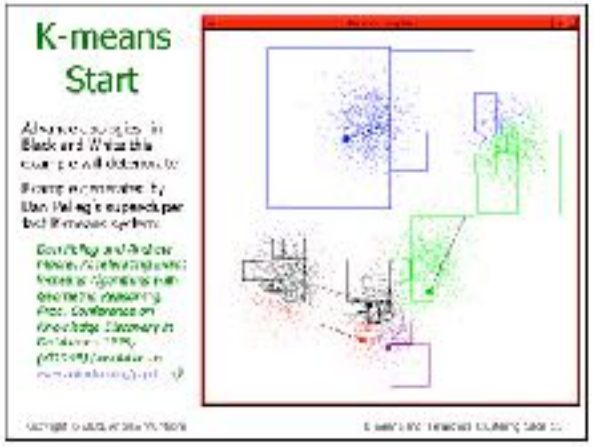


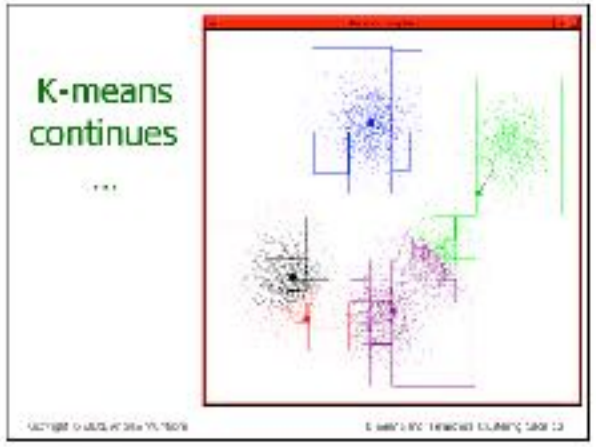


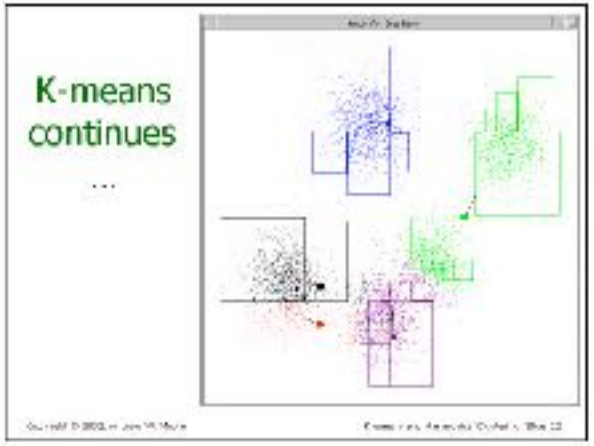
4

5

Example from Andrew Moore's tutorial on Clustering.



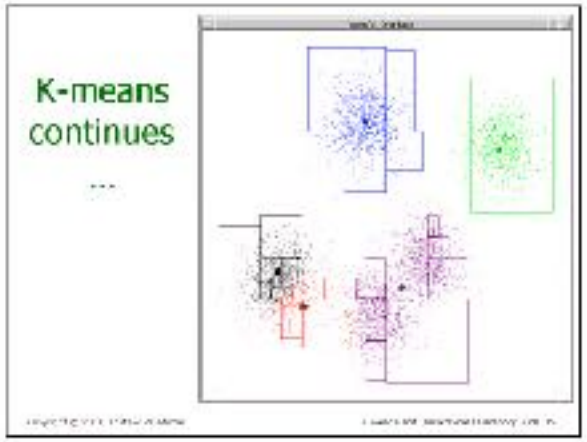


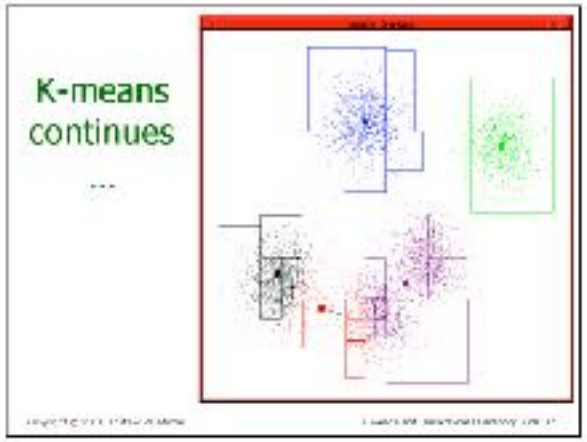


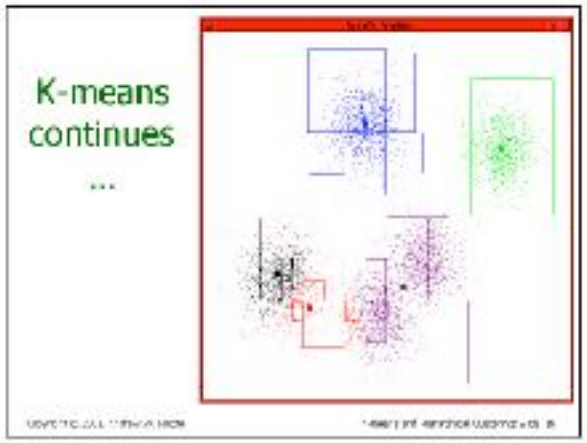
6

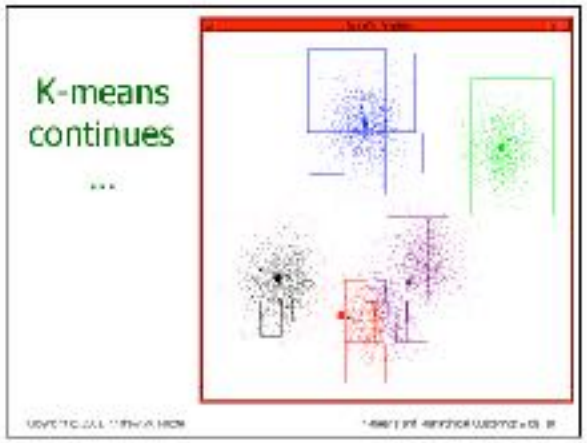
7

Example from Andrew Moore's tutorial on Clustering.







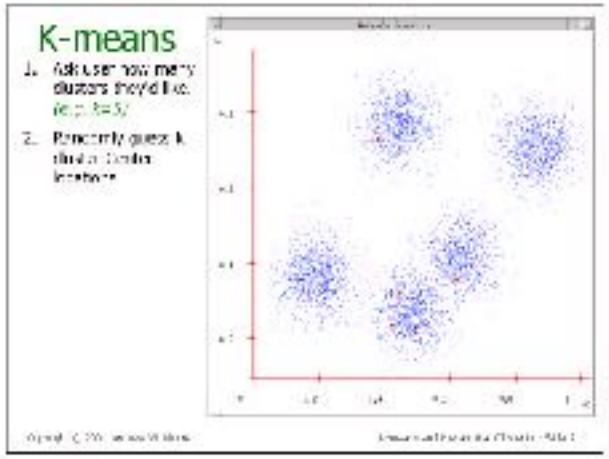


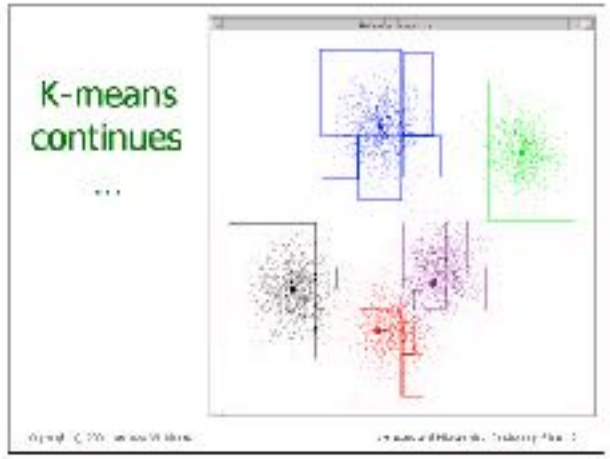
8

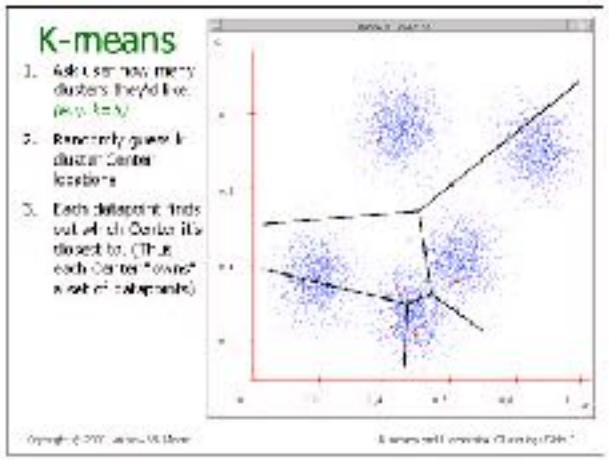
9

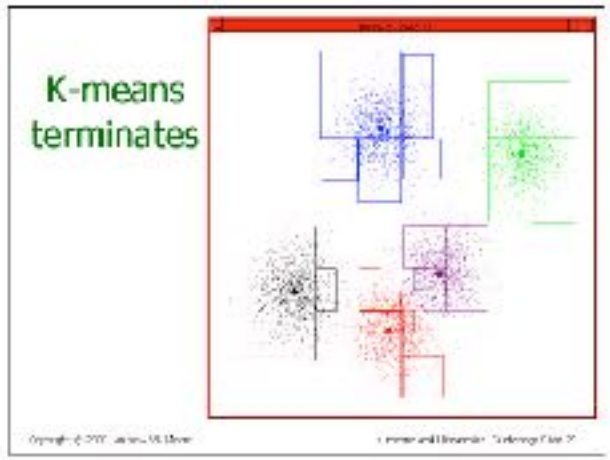
Example from Andrew Moore's tutorial on Clustering.

Start









End

4

10

Example from Andrew Moore's tutorial on Clustering.

K-Means Clustering [McQueen '67]

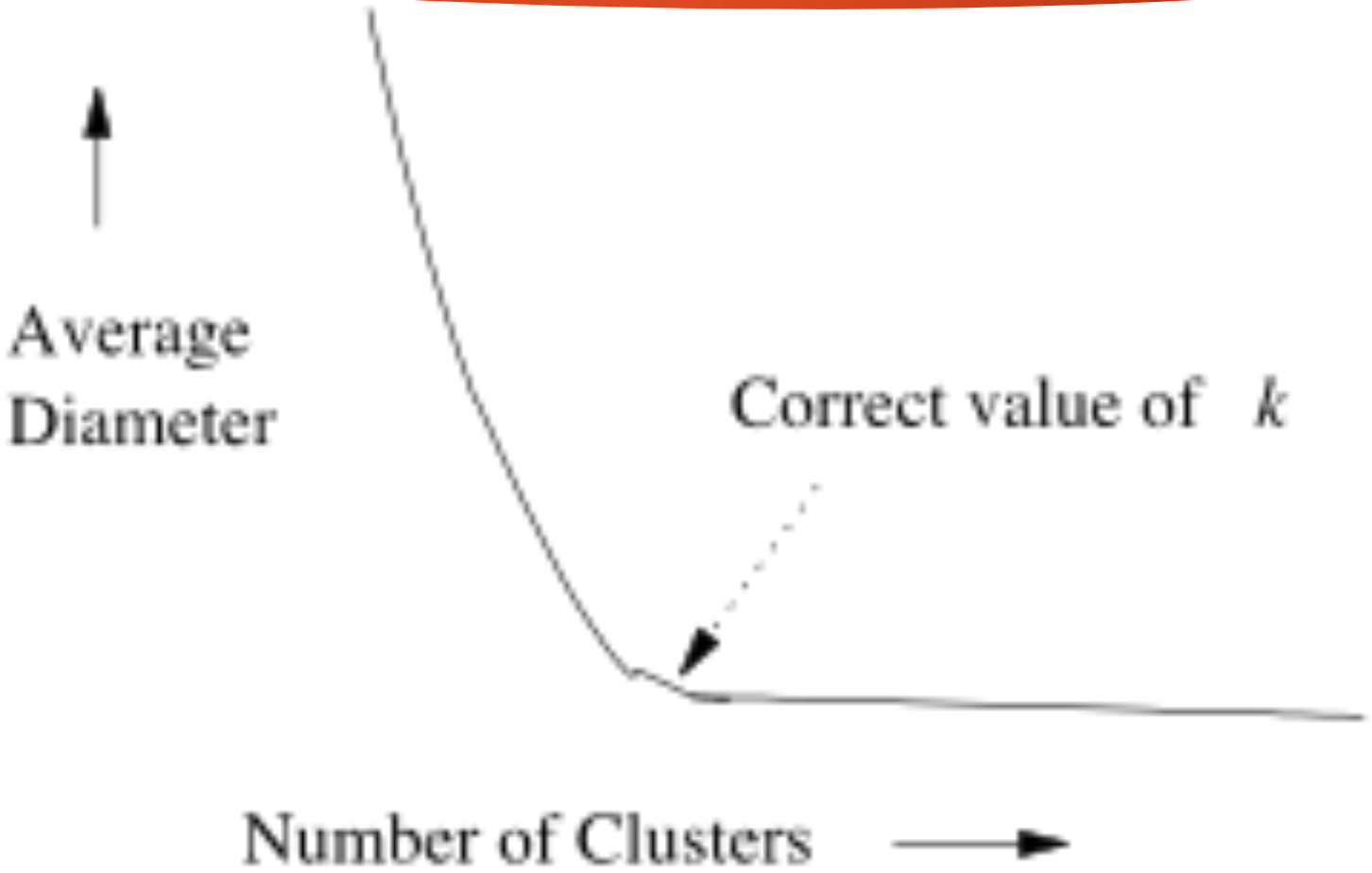
Repeat

- Start with randomly chosen cluster centers
- Assign points to give greatest increase in score
- Recompute cluster centers
- Reassign points

until (no changes)

Try the applet at: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

How to find K for K-means?



Comparisons

- Hierarchical clustering
  - Number of clusters not preset.
  - Complete hierarchy of clusters
  - Not very robust, not very efficient.
- K-Means
  - Need definition of a mean. Categorical data?
  - Can be sensitive to initial cluster centers; Stopping condition unclear
  - More efficient and often finds optimum clustering.