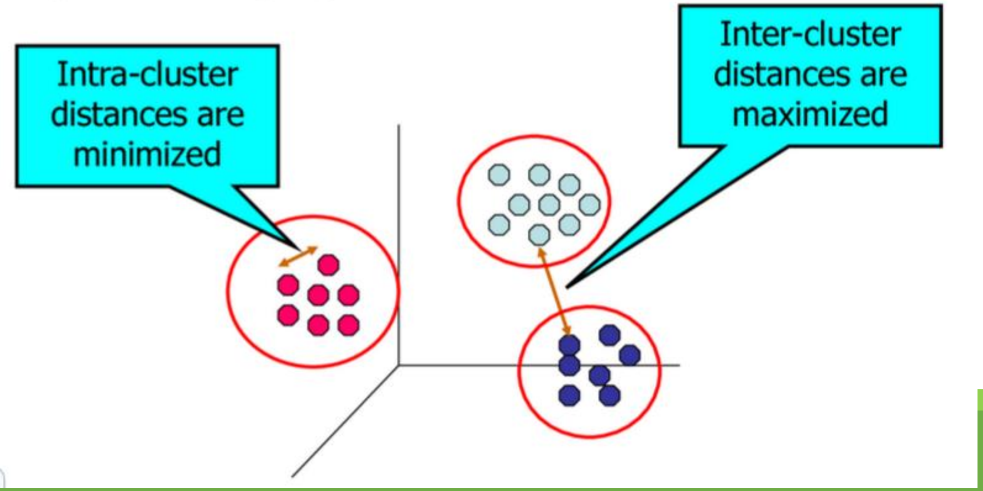
Cluster Analysis

Outline

- Clustering Basics
  - What is clustering, conceptually? What are the applications?
  - How to measure data similarity and distance?
- Different Approaches to Cluster Analysis
  - Hierarchical clustering
  - K-Means Clustering
  - Pre-processing: e.g., normalization and scaling
  - Output evaluation and post-processing

Clustering: The Main Idea

- Goal: Form groups (clusters) of similar records
- Segment data points into homogeneous (and, hopefully, meaningful) clusters
- Desired properties of clustering result:
  - High intra-cluster similarity, low inter-cluster similarity
- "Unsupervised" learning technique



Using Clustering

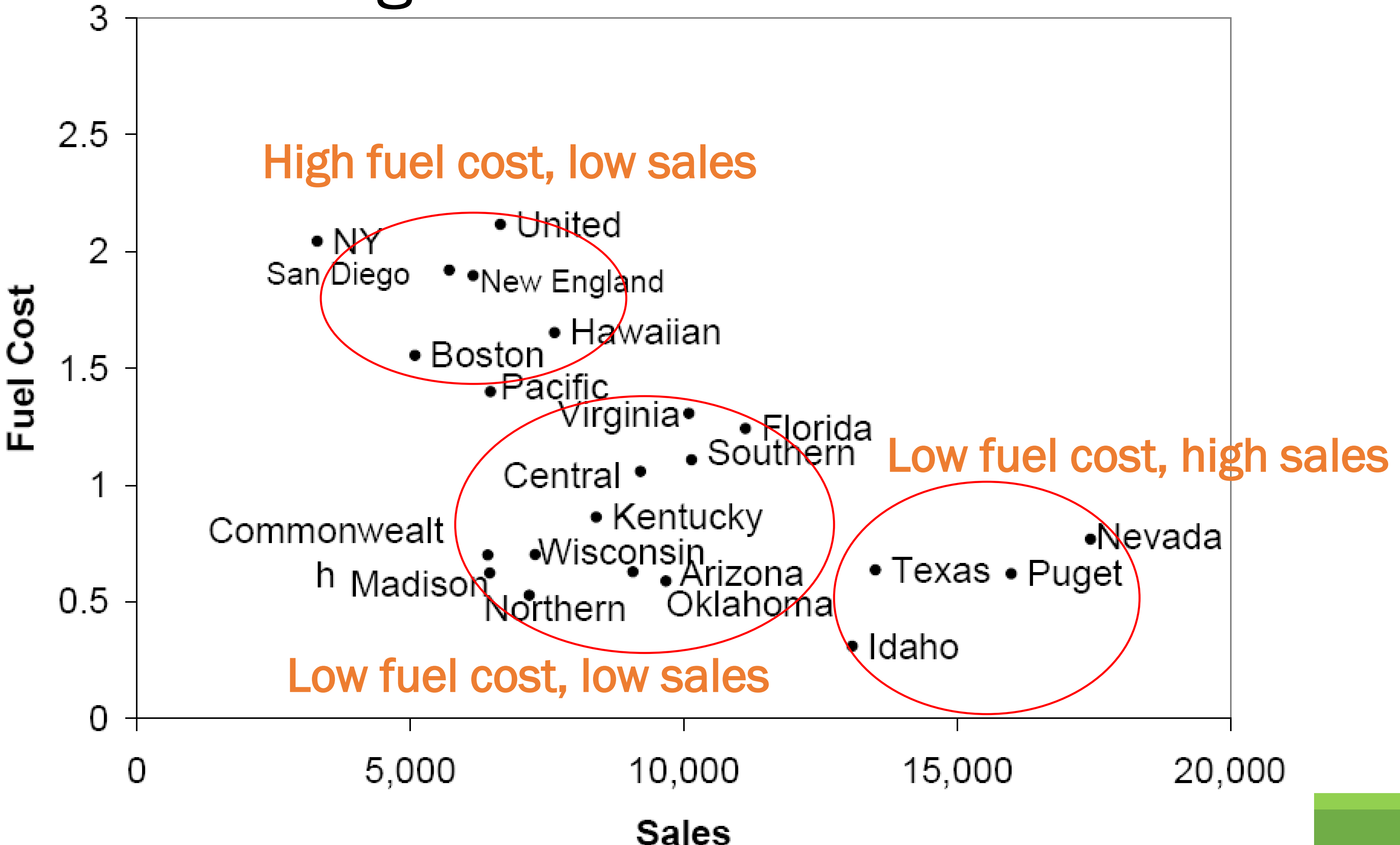
- Helps to gain insights into your data
  - Instead of trying to look at the entire dataset, you can inspect the representative clusters
- The basic idea has been used throughout the history
  - Periodic table of the elements
  - Classification of species
  - Grouping securities in portfolios
  - Grouping firms for structural analysis of economy
- Many applications
  - Market segmentation, medical diagnostics, bioinformatics, text mining / information retrieval, etc.

Example: Public Utilities

- **Goal:** based on the information about different public utility companies, find clusters/groups of similar utilities (according to their descriptive attributes)
- **Data:** 22 firms, 8 variables
  - Fixed-charge covering ratio
  - Rate of return on capital
  - Cost per kilowatt capacity
  - Annual load factor
  - Growth in peak demand
  - Sales
  - % nuclear
  - Fuel costs per kwh

Company	Fixed_charge	RoR	Cost	Load	Δ Demand	Sales	Nuclear	Fuel_Cost
Arizona	1.06	9.2	151	54.4	1.6	9077	0	0.628
Boston	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
Central	1.43	15.4	113	53	3.4	9212	0	1.058
Commonwealth	1.02	11.2	168	56	0.3	6423	34.3	0.7
Con Ed NY	1.49	8.8	192	51.2	1	3300	15.6	2.044
Florida	1.32	13.5	111	60	-2.2	11127	22.5	1.241
Hawaiian	1.22	12.2	175	67.6	2.2	7642	0	1.652
Idaho	1.1	9.2	245	57	3.3	13082	0	0.309
Kentucky	1.34	13	168	60.4	7.2	8406	0	0.862
Madison	1.12	12.4	197	53	2.7	6455	39.2	0.623
Nevada	0.75	7.5	173	51.5	6.5	17441	0	0.768
New England	1.13	10.9	178	62	3.7	6154	0	1.897
Northern	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
Oklahoma	1.09	12	96	49.8	1.4	9673	0	0.588
Pacific	0.96	7.6	164	62.2	-0.1	6468	0.9	1.4
Puget	1.16	9.9	252	56	9.2	15991	0	0.62
San Diego	0.76	6.4	136	61.9	9	5714	8.3	1.92
Southern	1.05	12.6	150	56.7	2.7	10140	0	1.108
Texas	1.16	11.7	104	54	-2.1	13507	0	0.636
Wisconsin	1.2	11.8	148	59.9	3.5	7287	41.1	0.702
United	1.04	8.6	204	61	3.5	6650	0	2.116
Virginia	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

Sales & Fuel Cost:
3 rough clusters can be seen



Extension to More Than 2 Dimensions

- In prior example, clustering was done by eye
  - Eyeballing only works for 2 or 3-dimensional data.

- Multiple dimensions require formal algorithm with
  - A **distance measure**
  - A way to use the distance measure in forming clusters

- Two types of algorithms: **hierarchical** and **non-hierarchical**

Two Popular Clustering Approaches

- Hierarchical clustering
  - Observations are sequentially grouped to create clusters
  - Based on distances between observations and distances between clusters

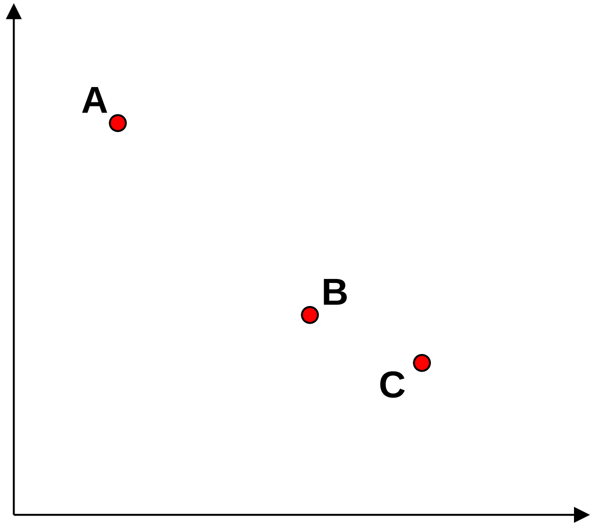
- K-means clustering (Nonhierarchical method)
  - Observations are allocated to one of a pre-specified set of clusters
  - Based on their distance from each cluster

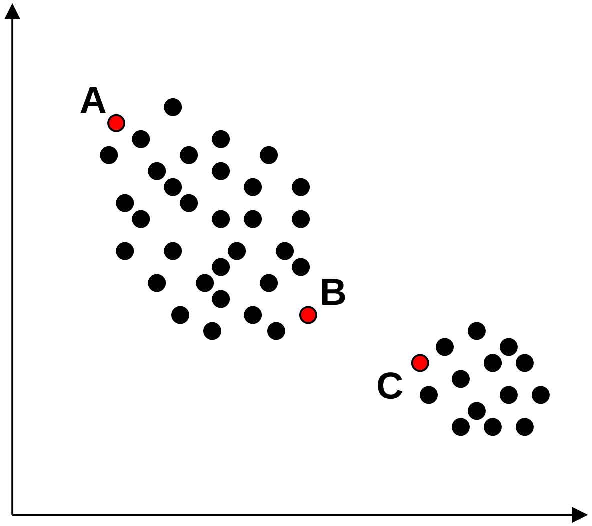
Things to Think About: General Issues

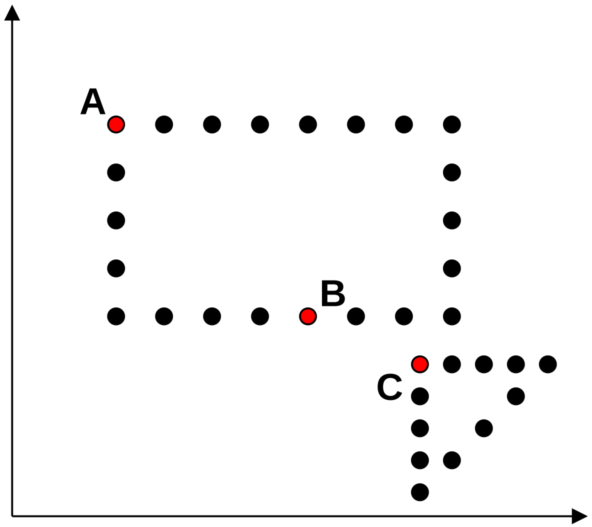
- Similarity metrics are important for clustering
  - Maximize intra-cluster similiarity (“tight” cluster)
  - Minimize inter-cluster similarity (“distinct” cluster)
  - But what kind of similarity to use?
- Stopping criteria
  - How many clusters should you have?
- Clustering algorithms and their parameters
  - How to come up the clusters?

What similarity metrics to use?

Consider data points A, B, and C. Is B more similar to A or to C?







Direct similarity
based on distance

Contextual similarity
Based not only on distance but the relationship with other points

Conceptual similarity
Based on whether they represent a common underlying concept

Most clustering techniques deal with these types of similarity

Distance Metric

- A key component of most clustering techniques is the **distance metric**
- Given the distance metric, clustering techniques will cluster data according to that metric
- Therefore, the choice of a metric is important
  - Distance between two points
  - Distance between clusters

Numerical distance Example

Euclidean distance:

d(M,N) = $\sqrt{(a_1 - b_1)^2 + \cdots + (a_k - b_k)^2}$

= c = $\sqrt{(a^2 + b^2)}$ = 13

Manhattan distance:

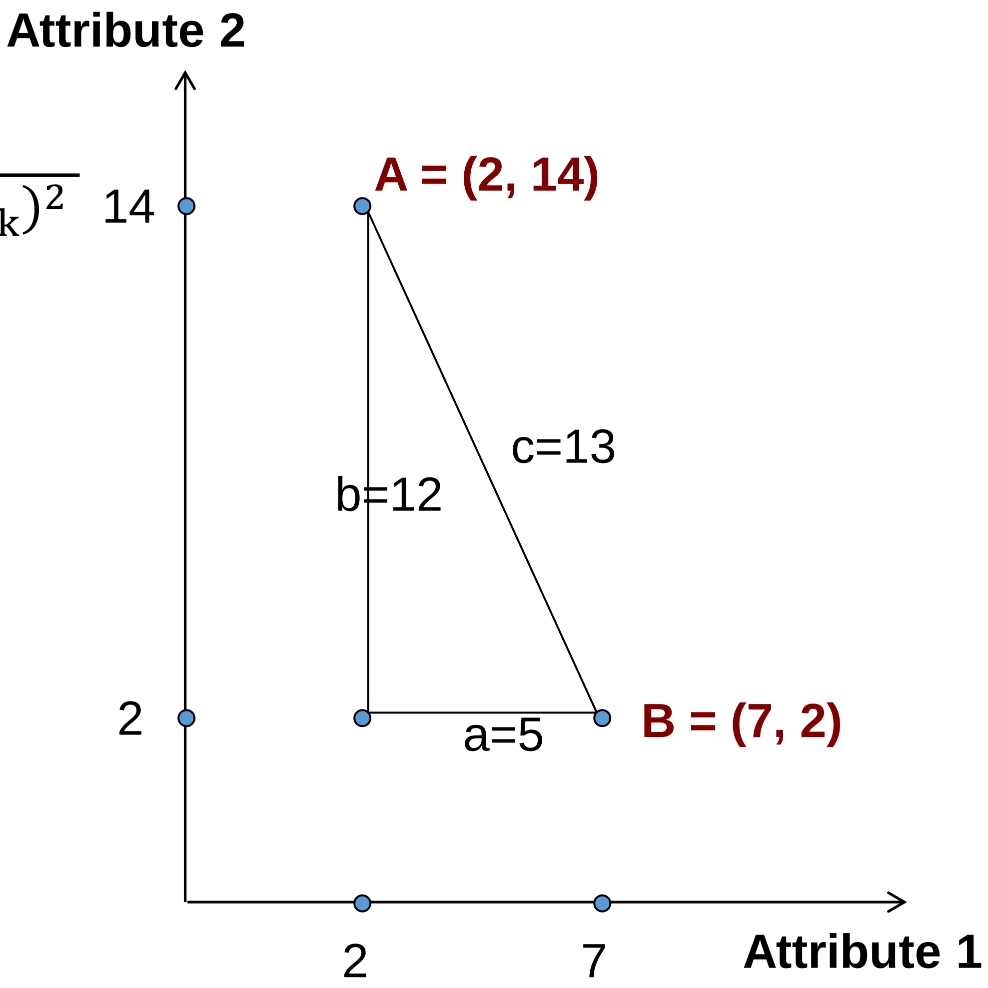
d(M,N) = $|a_1 - b_1| + \ldots + |a_k - b_k|$

= a + b = 17

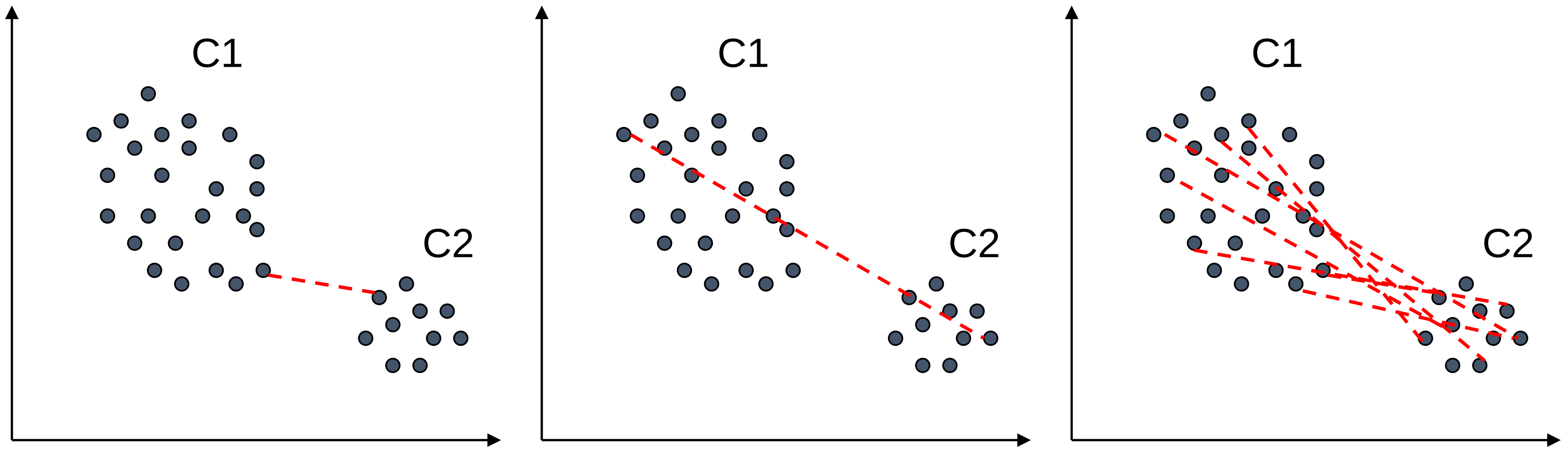
Max-coordinate distance:

d(M,N) = $\max_i |a_i - b_i|$

= max {a, b}= 12



Distance Between Clusters



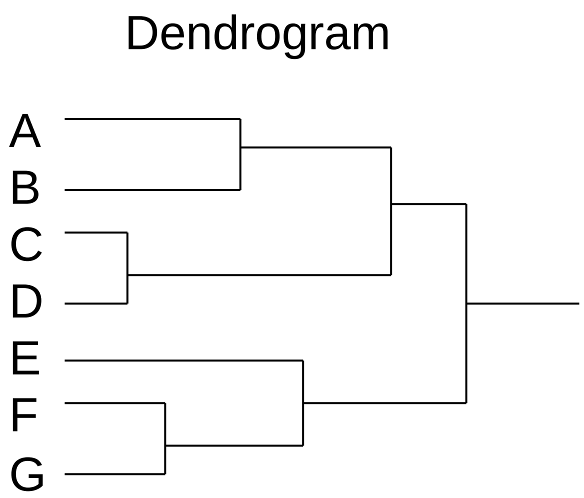
Single linkage
(minimum pairwise distance between points from two different clusters)

Complete linkage
(maximum pairwise distance between points from two different clusters)

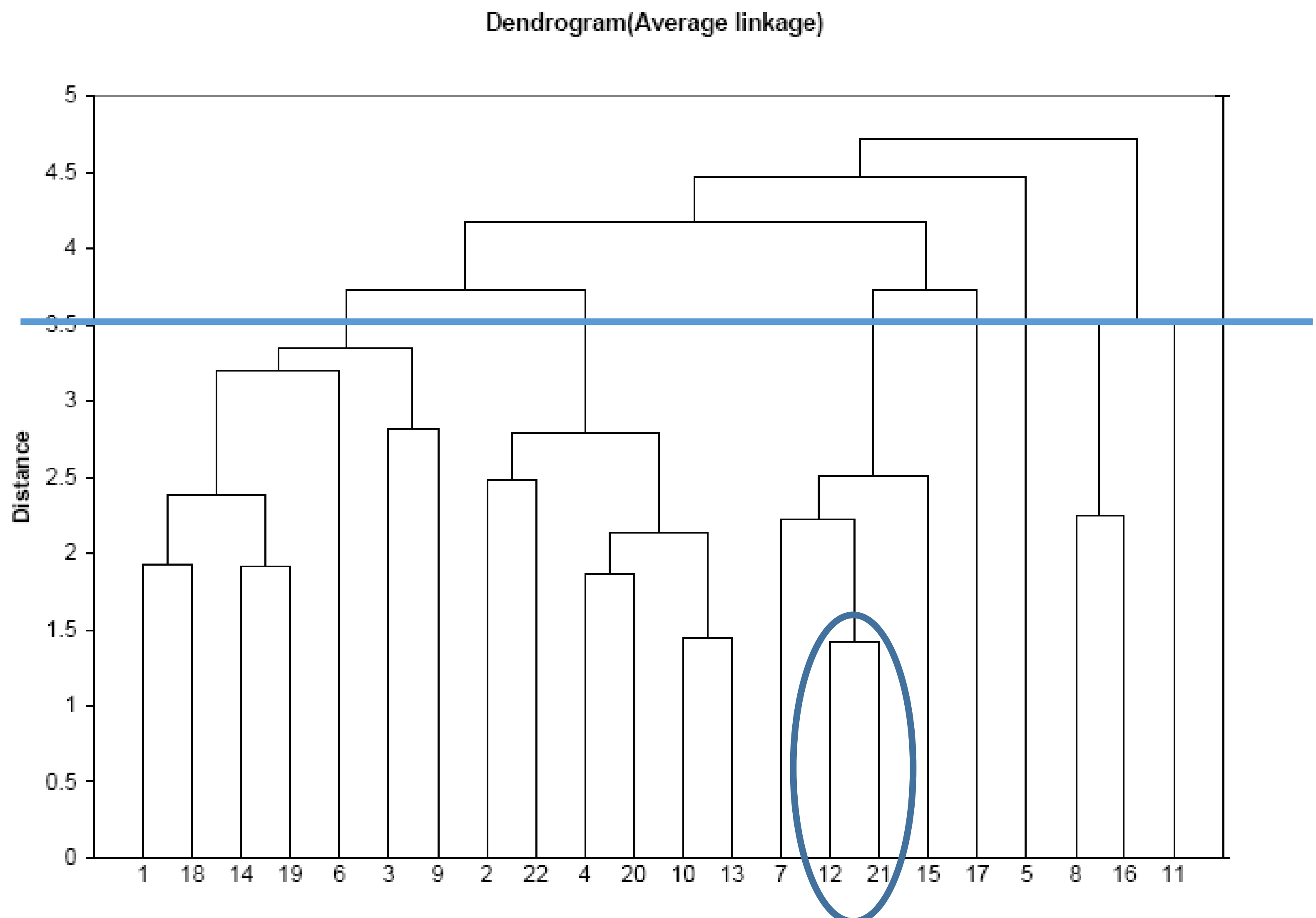
Average linkage
(average pairwise distance between points from two different clusters)

Hierarchical Methods

- Agglomerative (Bottom-Up)
  - Start with one cluster per data point
  - Merge two nearest clusters
    - criteria: min, max, avg distance
  - Repeat until only one cluster left
  - Output Dendrogram
- Divisive (top-down)
  - Start with one all-inclusive cluster
  - Repeatedly divide into smaller clusters
  - not as commonly used as agglomerative



Records 12 & 21 are closest & form first cluster



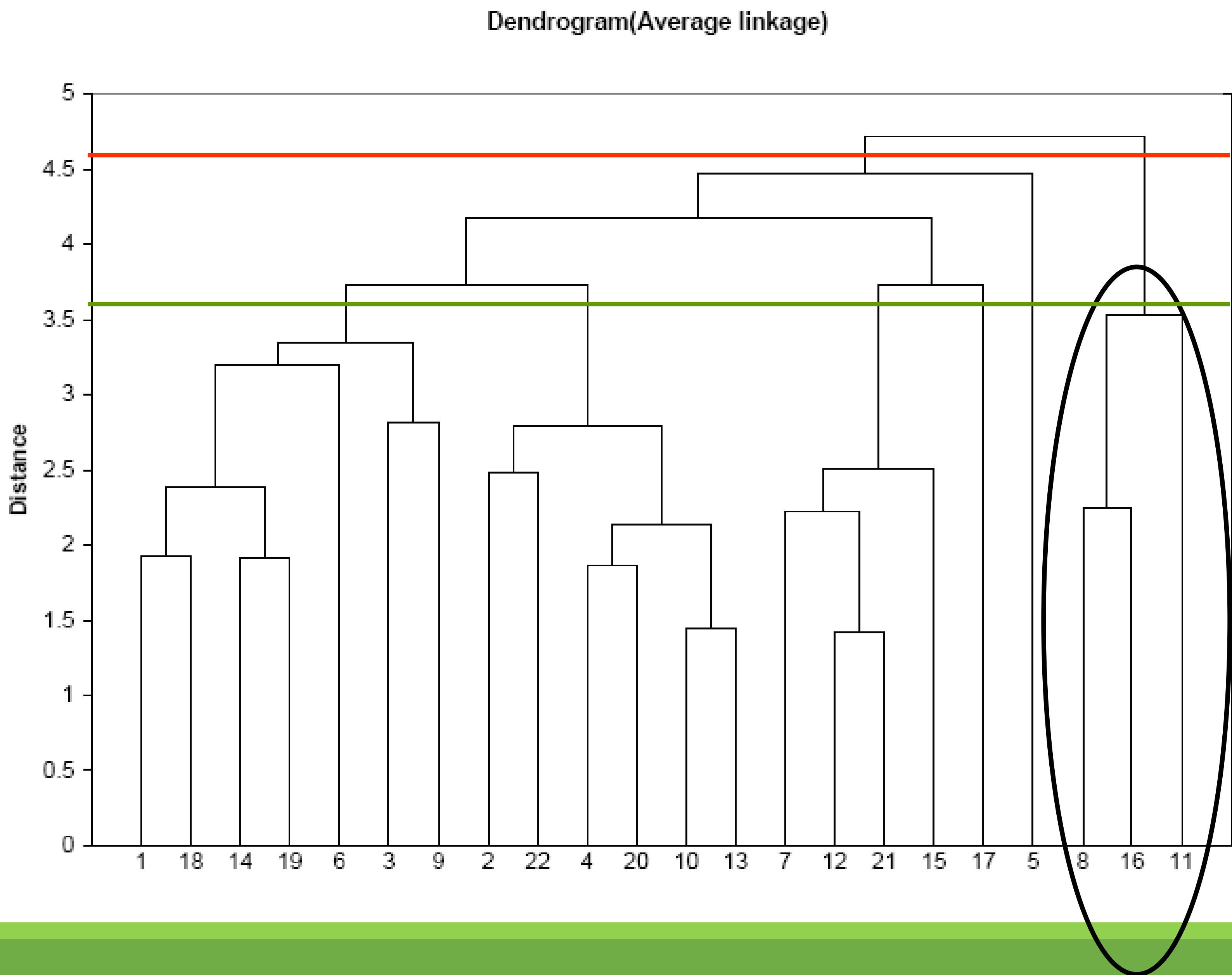
Reading the Dendrogram

See process of clustering: Lines connected lower down are merged earlier

- 10 and 13 will be merged next, after 12 & 21

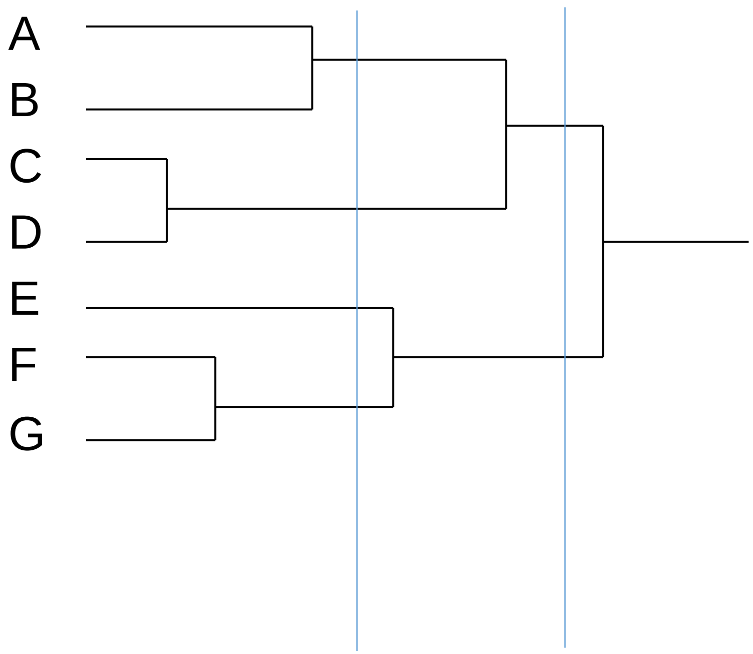
Determining number of clusters: For a given "distance between clusters", a horizontal line intersects the clusters, to create clusters

- E.g., at distance of 4.6 (**red line** in next slide), data can be reduced to 2 clusters -- The smaller of the two is circled
- At distance of 3.6 (**green line**) data can be reduced to 6 clusters, including the circled cluster



Dendrogram(Average linkage)
Distance
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
1 18 14 19 6 3 9 2 22 4 20 10 13 7 12 21 15 17 5 8 16 11

Dendrogram



Questions:

- Are C&D more similar to each other than A&B?
- What is the best 2-cluster solution?
- What is the best 4-cluster solution?

Agglomerative Clustering

- The hierarchical approach used by XLMiner
- Advantages
  - Does not require prespecify number of clusters
  - Dendrogram represents the cluster process and results in an intuitive way
- Limitations
  - Slow & computational expensive
    - Need to compute *n* x *n* distance matrix
  - Low stability
    - Easily affected by reordering data or dropped cases
  - Sensitive to outliers

Validating Clusters

Interpretation

Goal: obtain meaningful and useful clusters

Caveats:

(1) Random chance can often produce apparent clusters

(2) Different cluster methods produce different results

Solutions:

- Obtain summary statistics
- Also review clusters in terms of variables **not** used in clustering
- Label the cluster (e.g. clustering of financial firms in 2008 might yield label like "midsize, sub-prime loser")

Desirable Cluster Features

- **Stability** – are clusters and cluster assignments sensitive to slight changes in inputs? Are cluster assignments in partition B similar to partition A?

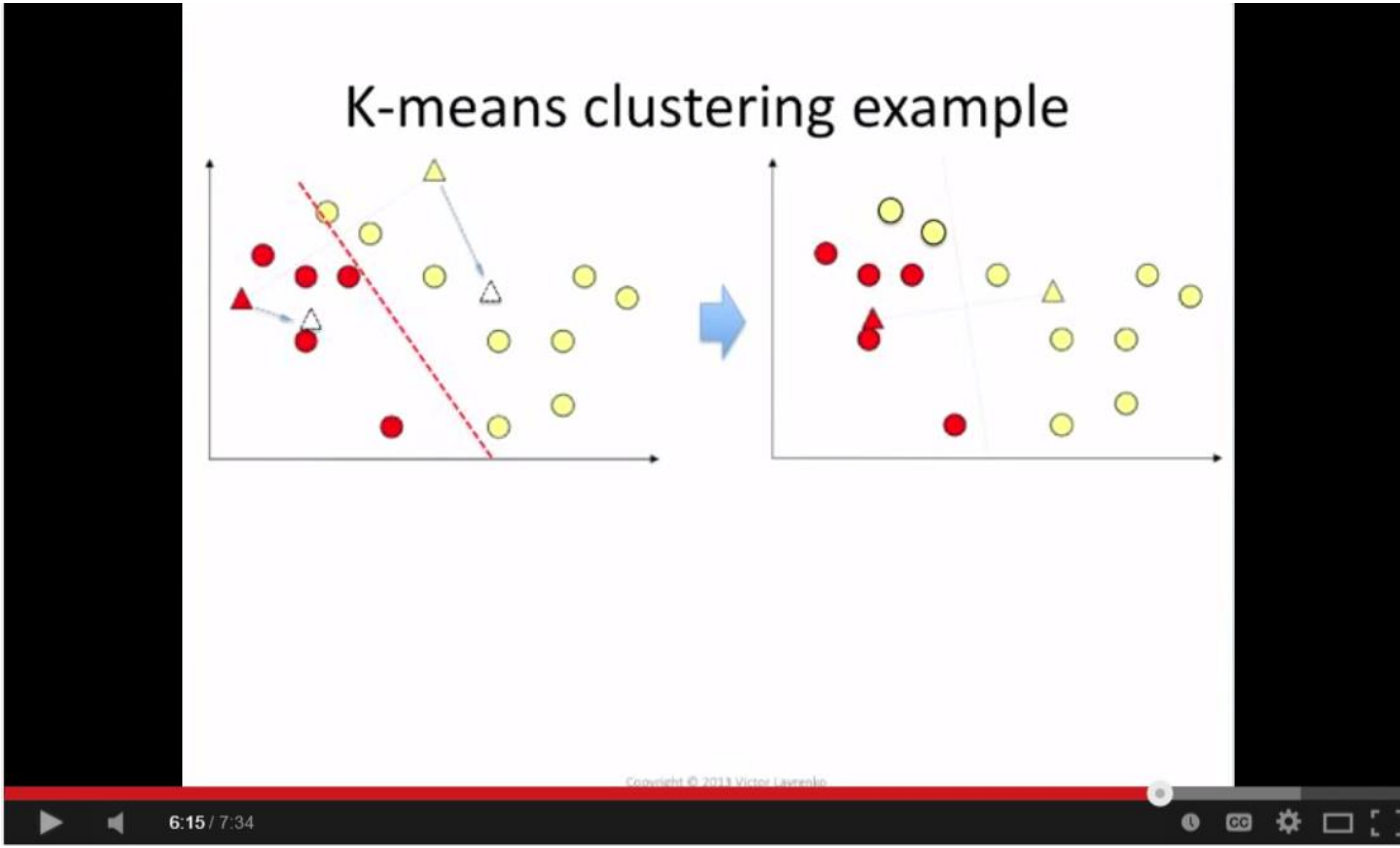
- **Separation** – check ratio of *between-cluster* variation to *within-cluster* variation (higher is better)

Nonhierarchical Clustering:
K-Means Clustering

K-Means Clustering Algorithm

1. Choose # of clusters desired, *k*
2. Start with a partition into k clusters
   Often based on random selection of k centroids
3. At each step, move each record to cluster with closest centroid
4. Recompute centroids, repeat step 3
5. Stop when moving records increases within-cluster dispersion

K-means Demo



Pre-process to get good centroids

- Get the data ready for analysis
  - Deal with missing values
  - Address any measurement error
- Rescale (e.g., Normalize) the Data
  - Reduces dispersion of data points by re-computing the distance
  - Preserves differences while dampening the effect of the outliers
- Remove Outliers
  - Reduces dispersion of data points by removing atypical data
  - They don’t represent the population anyway
  - Big field of study now in data mining (has applications for fraud detection, discovery of blockbuster drugs in pharmaceuticals)

Rescaling

- **Problem**: Raw distance measures are highly influenced by scale of measurements
  - E.g., income=$100,000; height=1.50m (income will highly dominate height in distance computations)

- Solution: normalize/standardize the data first
  - Rationale: transform variables into similar scale so that they can be compared and can contribute equally to the distance computations

Standardization and Normalization

- **Standardization** adjusts the intervals of attributes to a common range (also known as **min-max scaling**)
  - Calculate standardized values in interval [0,1]

$$\mathbf{q = (x - min) / (max - min)}$$

    - x – original value of the attribute
    - min/max – smallest/largest value of the attribute
    - q – resulting (scaled) value of the attribute in the range [0,1]
- **Normalization** (**z-score**) shift values to a normal curve with mean 0 and variance 1.

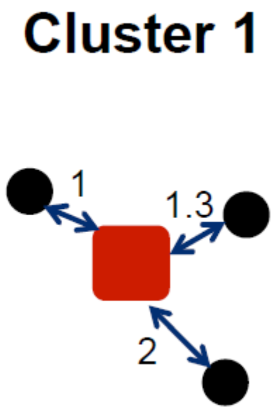
$$\mathbf{z = (x - m) / s}$$

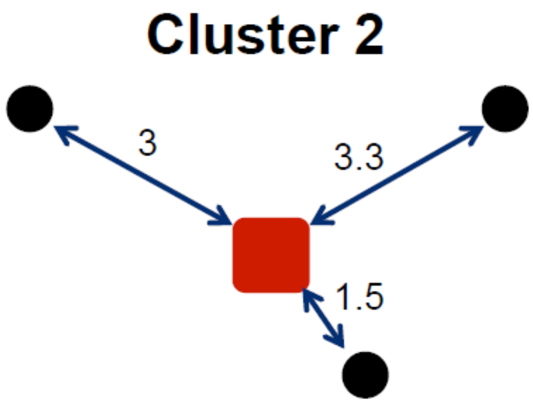
    - m – mean value of the attribute
    - s – standard deviation (or mean absolute deviation)

Evaluating K-means Cluster Results

- Sum-of-Squares Error (SSE)
  - The distance to the nearest cluster center
  - How close does each point get to the center?
  - $SSE = \sum_{i=1}^{K} \sum_{x \in C_i} d(m_i, x)^2$
- This just means:
  - In a cluster *i*, compute distance from a point *x* to the cluster center m_i.
  - Square the distance (so sign is not an issue)
  - Add them all together.

Example: Evaluating Clusters





$SSE_1 = 1^2 + 1.3^2 + 2^2$

$= 1 + 1.69 + 4$

$= 6.69$

$SSE_2 = 3^2 + 3.3^2 + 1.5^2$

$= 9 + 10.89 + 2.25$

$= 22.14$

Lower individual cluster SSE = a better cluster (more **Cohesion**)
Lower total SSE = a better set of clusters
More clusters?

Choosing Best Initial Centroids

- There is no single, best way to choose initial centroids
- So what do you do?
  - Multiple runs
  - Use a subsample first and then apply it to your main data set
  - Select more centroids to start with, then choose the ones that are farthest apart (most distinct)

Post-Processing: Better Centroids

- "Post": Interpreting the results of the cluster Analysis
- Remove small clusters
  - May be outliers
- Split loose clusters
  - With high SSE that look like they are really two different groups
- Merge clusters
  - With relatively low SSE that are "close" together

Distance Between Clusters

Distance between	Cluster-1	Cluster-2	Cluster-3
Cluster-1	0	5.03216253	3.16901457
Cluster-2	5.03216253	0	3.76581196
Cluster-3	3.16901457	3.76581196	0

Clusters 1 and 2 are relatively well-separated from each other, while cluster 3 not as much

Within-Cluster Dispersion

Data summary (In Original coordinates)

Cluster	#Obs	Average distance in cluster
Cluster-1	12	1748.348058
Cluster-2	3	907.6919822
Cluster-3	7	3625.242085
Overall	22	2230.906692

Clusters 1 and 2 are relatively tight, cluster 3 very loose

Conclusion: Clusters 1 & 2 well defined, not so for cluster 3

Next step: try again with *k*=2 or *k*=4

Limitations of k-Means Clustering

- K-Means gives unreliable results when…
  - Clusters vary widely in size
  - Clusters vary widely in density
  - Clusters are not in rounded shapes
  - The data set has a lot of outliers

- In these cases
  - The clusters may never make sense
  - Either the data is not suitable for clustering
  - Or different clustering technique should be used.

Cluster Analysis: Summary

- Cluster analysis is an exploratory tool
  - Useful when it produces meaningful clusters
- **Hierarchical** clustering
  - gives visual representation of different levels of clustering
  - Can vary highly depending on settings
  - Computationally expensive
- Centroid-based clustering (**k-means**)
  - computationally cheaper and more stable
  - requires user to set k
- Can use both methods
- Be wary of chance results; data may not have definitive "real" clusters