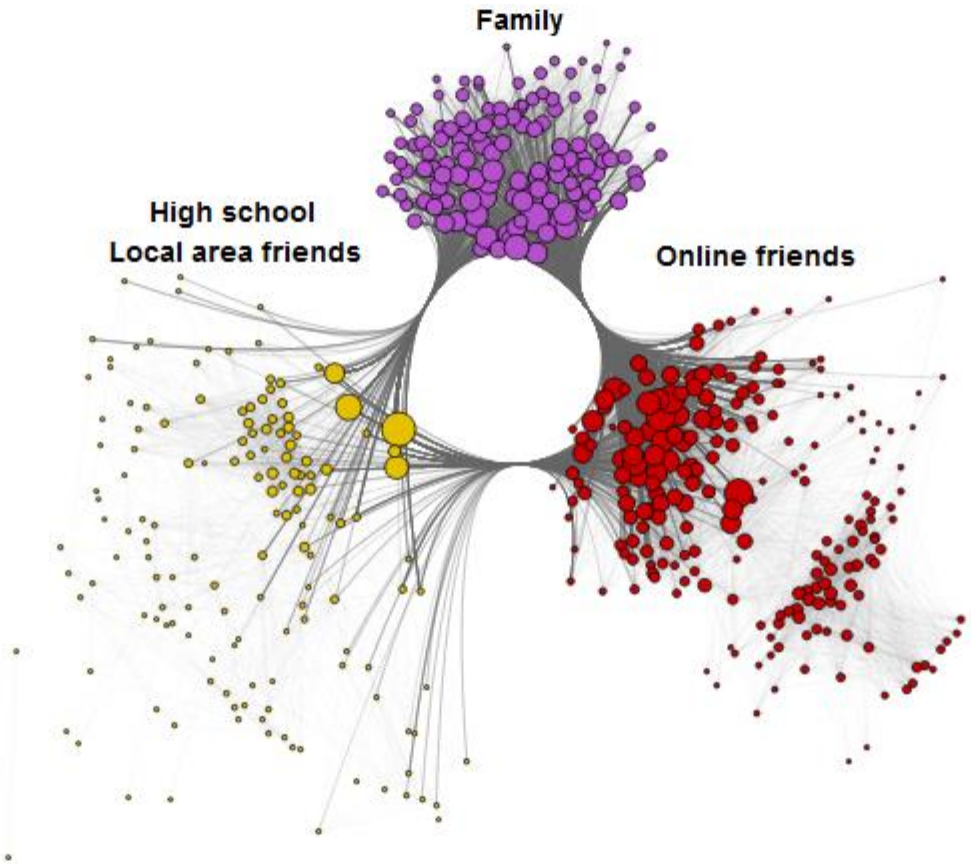
Clustering



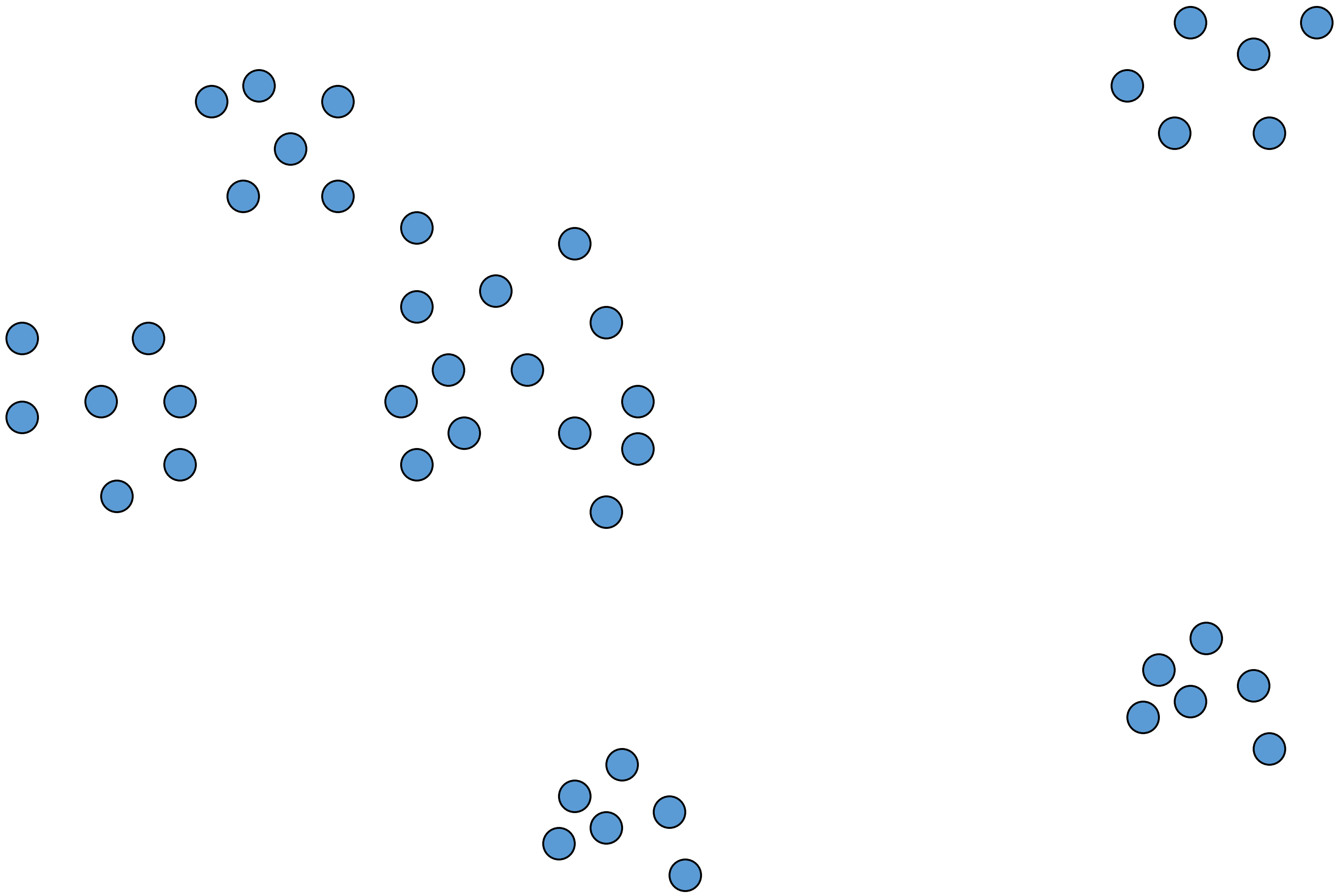
Clustering vs. Class prediction

- Class prediction:
  - A *learning set* of objects with known classes
  - Goal: put new objects into existing classes
  - Also called: *Supervised learning,* or *classification*
- Clustering:
  - No learning set, no given classes
  - Goal: discover the ”best” classes or groupings
  - Also called: *Unsupervised learning*, or *class discovery*

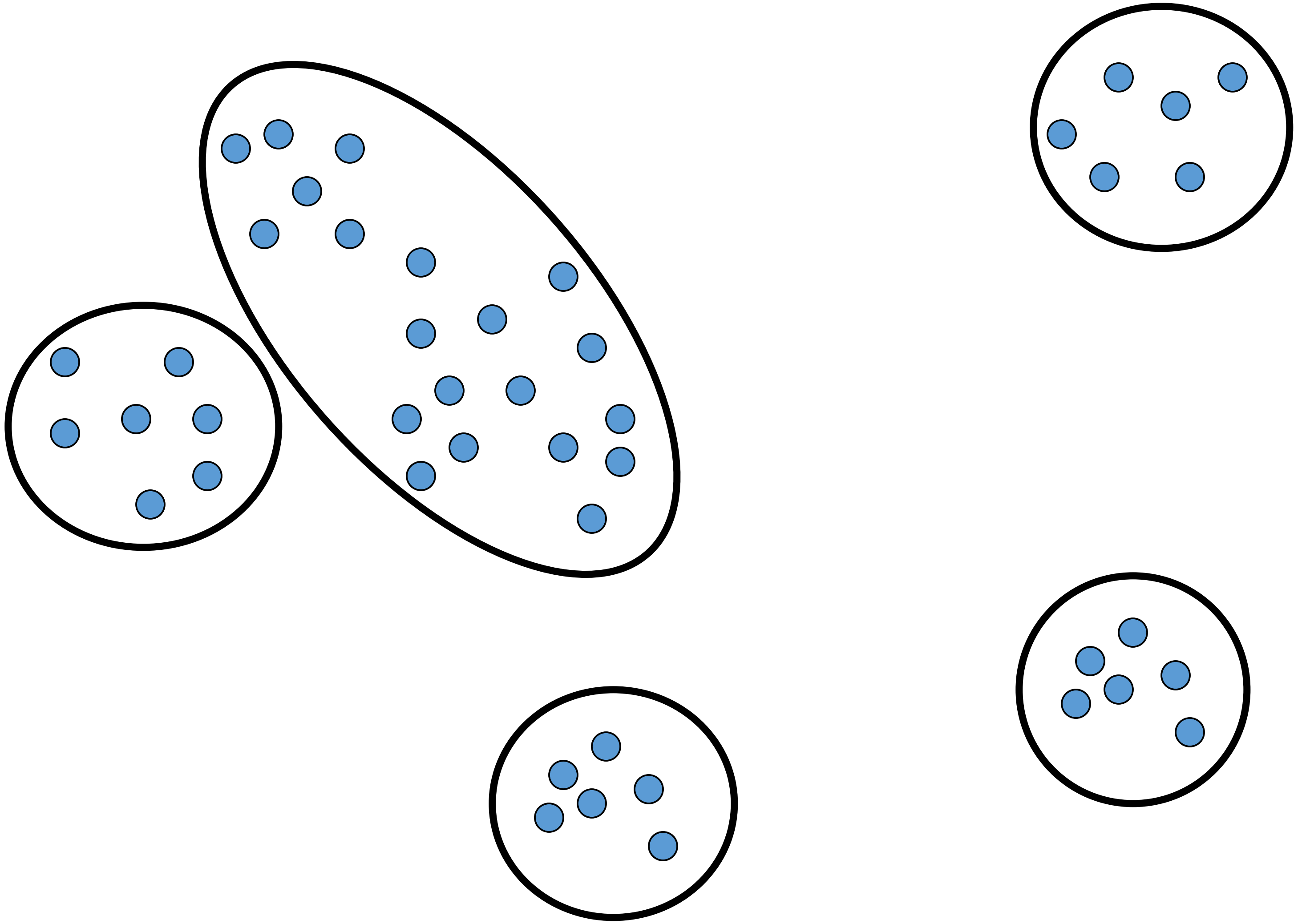
Clustering

- Clustering: the process of grouping a set of objects into classes of similar objects
- Most common form of *unsupervised learning*
    - Unsupervised learning = learning from raw data, as opposed to supervised data where a classification of examples is given

Clustering



Clustering



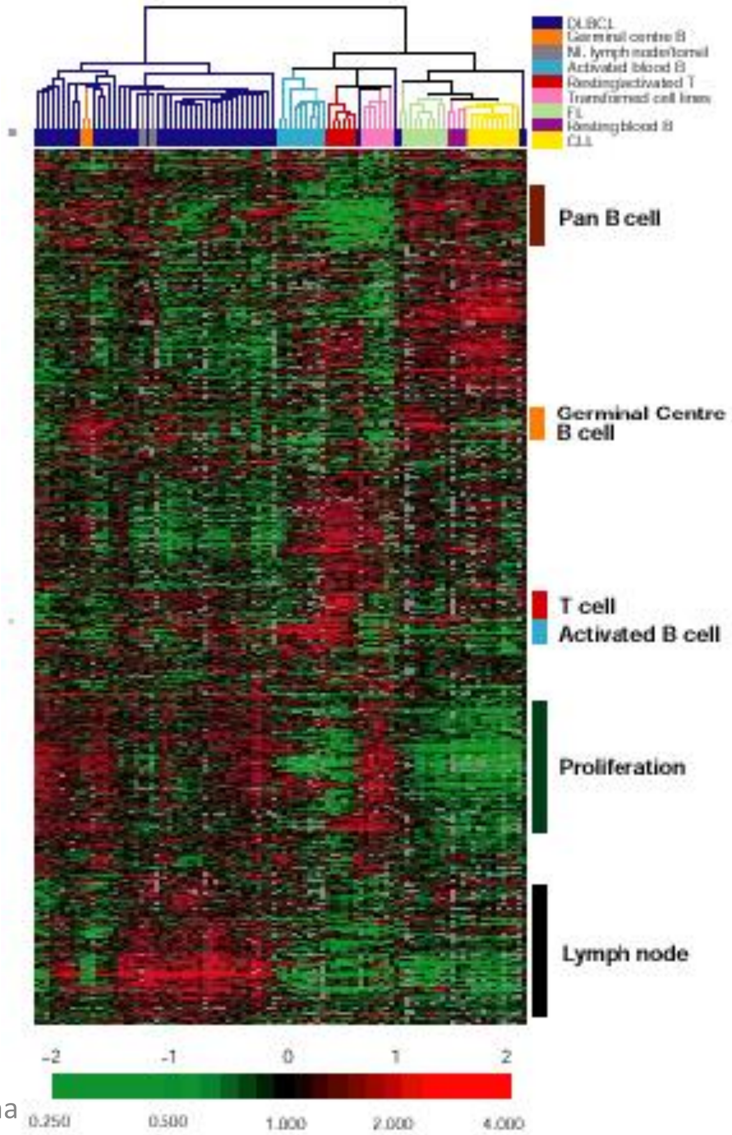
Issues in clustering

- Used to *explore* and *visualize* data, with few preconceptions
- Many subjective choices must be made, so a clustering output tends to be subjective
- It is difficult to get truly statistically ”significant” conclusions
- Algorithms will always produce clusters, whether any exist in the data or not

Clustering

- Cluster or Classify genes according to tumors

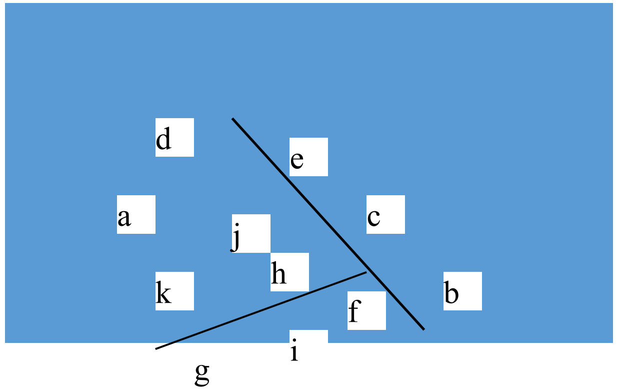
- Cluster tumors according to genes



Clusters: exclusive vs. overlapping

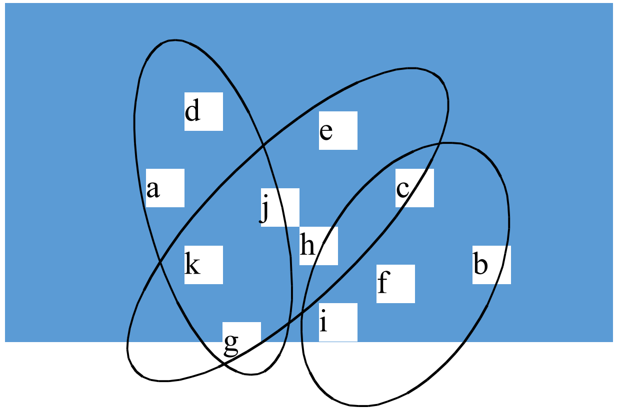
Simple 2-D representation

Non-overlapping



Venn diagram

Overlapping



Clustering considerations

- What does it mean for objects to be similar?
- What algorithm and approach do we take?
  - Top-down: k-means
  - Bottom-up: hierarchical agglomerative clustering
- Do we need a hierarchical arrangement of clusters?
- How many clusters?
- Can we label or name the clusters?
- How do we make it efficient and scalable?

Steps in clustering

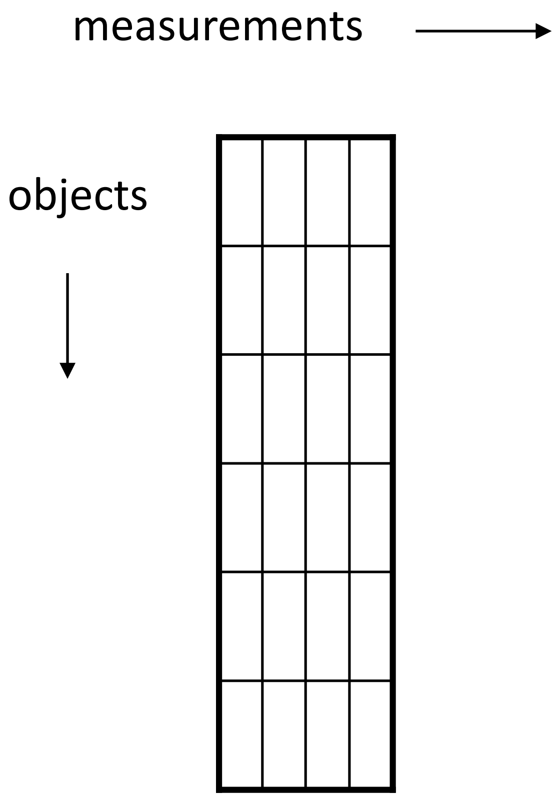
1. Feature selection and extraction
2. Defining and computing similarities
3. Clustering or grouping objects
4. Assessing, presenting, and using the result

Feature selection and extraction

- Deciding which measurements matter for similarity
- Data reduction
- Filtering away objects
- Normalization of measurements

The data matrix

- Every row contains the measurements for one object.
- Similarities are computed between all pairs of rows



Defining and computing similarities

- *Want clusters of instances that are similar to each other but dissimilar to others*
- Need a similarity measure
- Continuous case
  - Euclidean measure (compact isolated clusters)
  - The squared Mahalanobis distance
$$d_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)\Sigma^{-1}(\mathbf{x}_i - \mathbf{x}_j)^T$$
    alleviates problems with correlation
  - Many more measures

Defining and computing similarities: Euclidian Distance

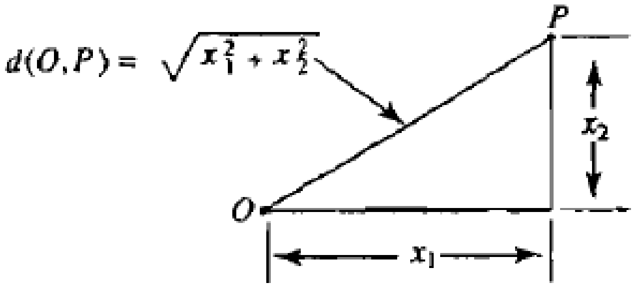


Figure 1.19 Distance given by the Pythagorean theorem.

$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

$$d(O,P) = \sqrt{x_1^2 + x_2^2}$$

$$d(O,P) = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$$

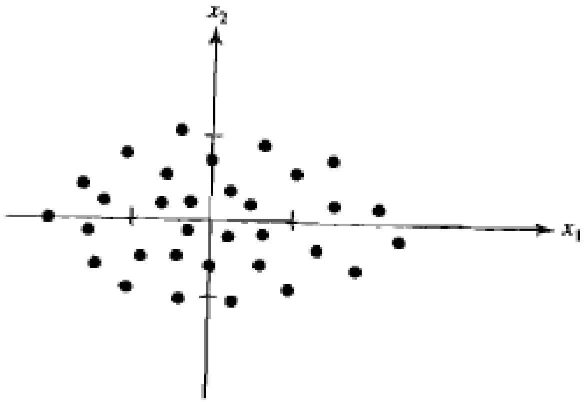


Figure 1.20 A scatter plot with greater variability in the x_1 direction than in the x_2 direction.

Defining and computing similarities: Mahalanobis (Statistical) Distance

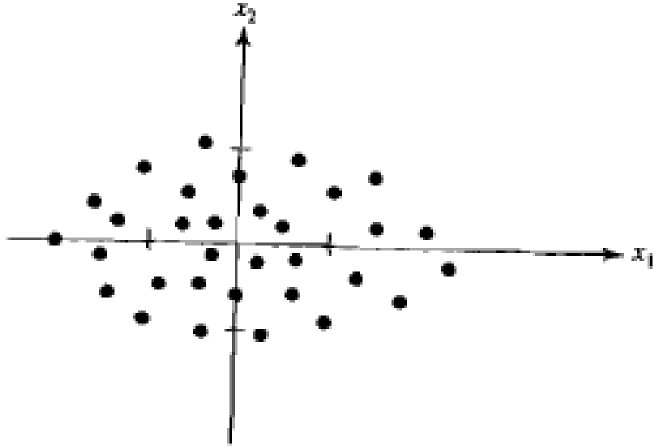
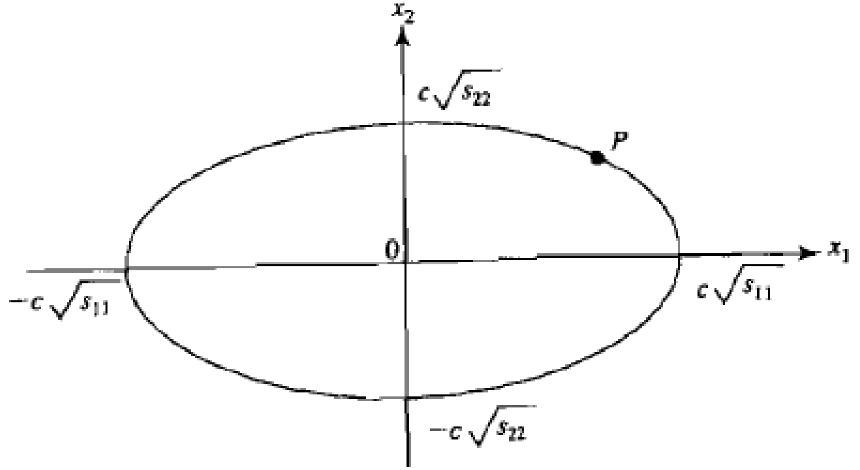


Figure 1.20 A scatter plot with greater variability in the x_1 direction than in the x_2 direction.

$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}}}$$

$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}} + \cdots + \frac{(x_p - y_p)^2}{s_{pp}}}$$



$$d(O,P) = \sqrt{(x_1^*)^2 + (x_2^*)^2}$$

$$= \sqrt{\left(\frac{x_1}{\sqrt{s_{11}}}\right)^2 + \left(\frac{x_2}{\sqrt{s_{22}}}\right)^2} = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}}$$

$$\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} = c^2$$

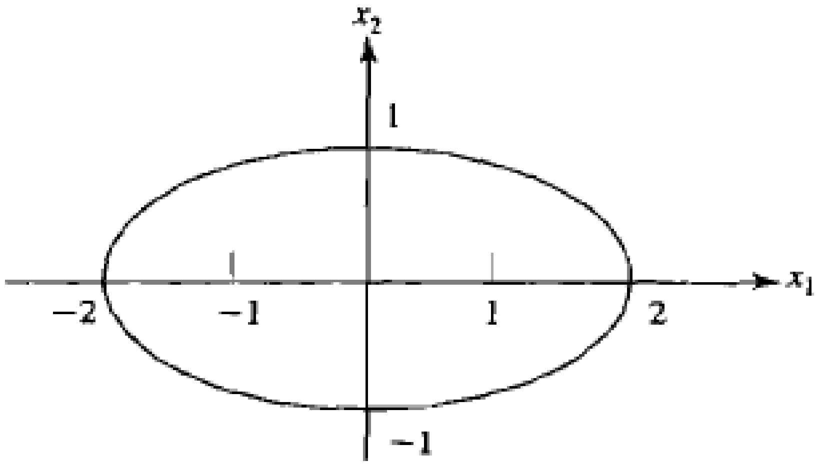


Figure 1.22 Ellipse of unit distance, $\frac{x_1^2}{4} + \frac{x_2^2}{1} = 1$.

Coordinates: (x_1, x_2)	Distance: $\frac{x_1^2}{4} + \frac{x_2^2}{1} = 1$
$(0, 1)$	$\frac{0^2}{4} + \frac{1^2}{1} = 1$
$(0, -1)$	$\frac{0^2}{4} + \frac{(-1)^2}{1} = 1$
$(2, 0)$	$\frac{2^2}{4} + \frac{0^2}{1} = 1$
$(1, \sqrt{3}/2)$	$\frac{1^2}{4} + \frac{(\sqrt{3}/2)^2}{1} = 1$

Mahalanobis Distance

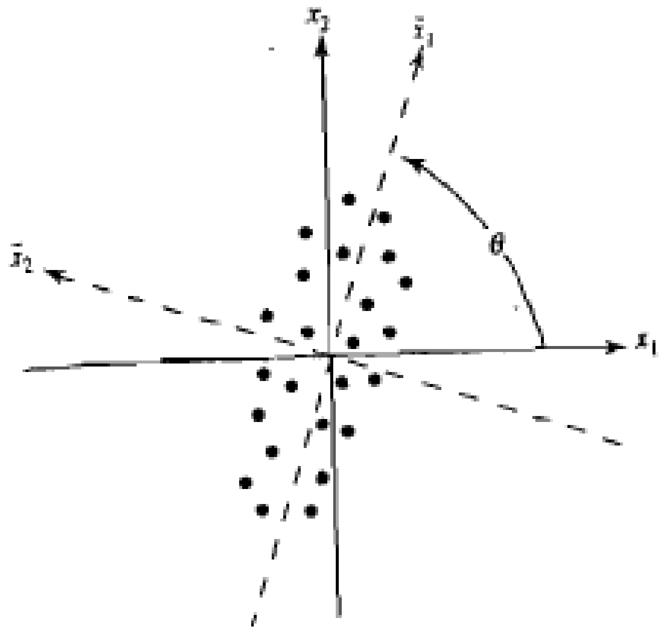


Figure 1.23 A scatter plot for positively correlated measurements and a rotated coordinate system.

$$d(O, P) = \sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}}}$$

$$\tilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta)$$

$$\tilde{x}_2 = -x_1 \sin(\theta) + x_2 \cos(\theta)$$

Defining and computing similarities

- Nominal attributes

$$d(\mathbf{x}_i, \mathbf{x}_j) = \frac{n - x}{n}$$

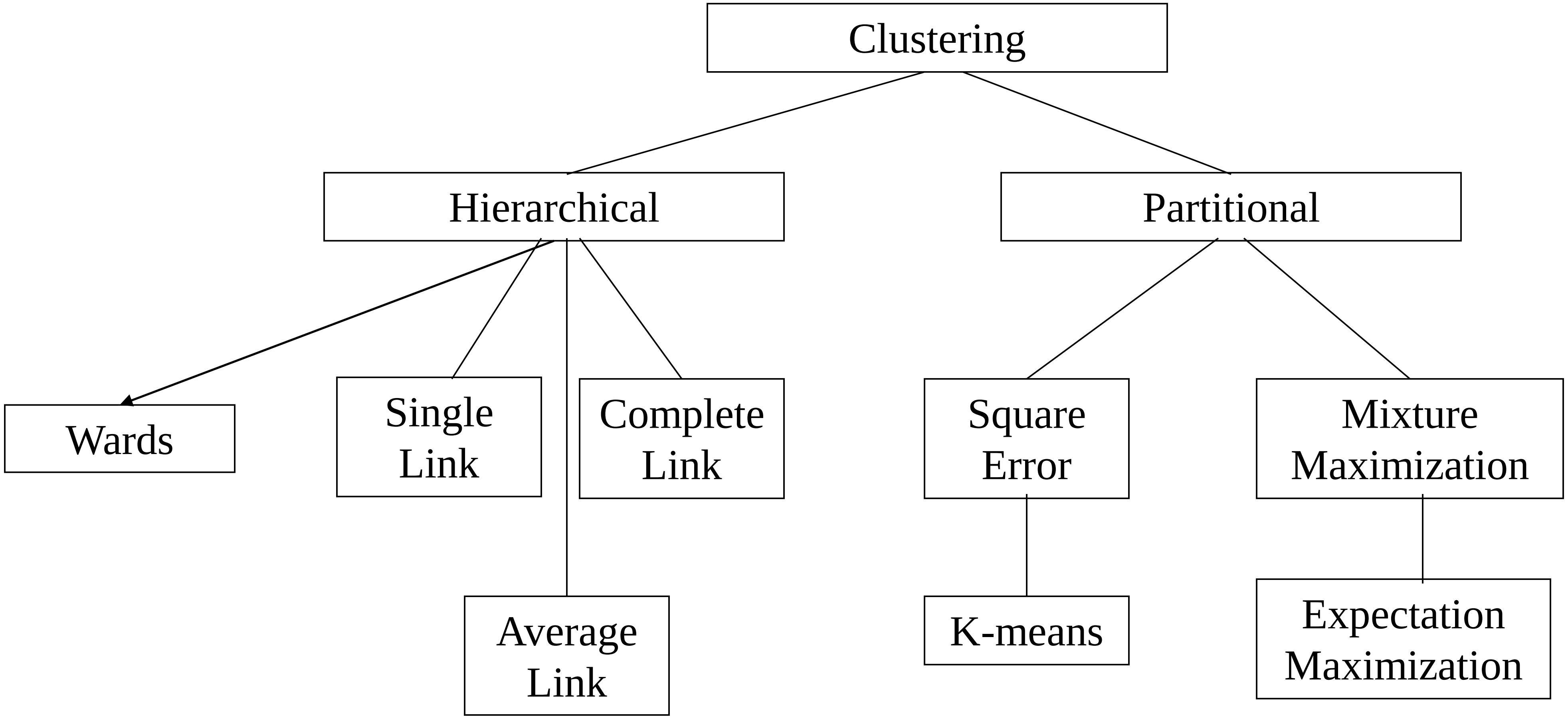
$$n = \text{Number of attributes}$$

$$x = \text{Number of attributes that are the same}$$

Clustering or grouping

- Hierarchical clusterings
  - Divisive: Starts with one big cluster and subdivides on cluster in each step
  - Agglomerative: Starts with each object in separate cluster. In each step, joins the two closest clusters
- Partitional clusterings
- Probabilistic or fuzzy clusterings

Clustering Techniques



Technique Characteristics

- Agglomerative vs Divisive
  - *Agglomerative*: each instance is its own cluster and the algorithm merges clusters
  - *Divisive*: begins with all instances in one cluster and divides it up
- Hard vs Fuzzy
  - Hard clustering assigns each instance to one cluster whereas in fuzzy clustering assigns degree of membership

More Characteristics

- Monothetic vs Polythetic
  - *Polythetic*: all attributes are used simultaneously, e.g., to calculate distance (most algorithms)
  - *Monothetic*: attributes are considered one at a time
- Incremental vs Non-Incremental
  - With large data sets it may be necessary to consider only part of the data at a time (data mining)
  - Incremental works instance by instance

Partitional Clustering

Partitional Clustering

- Output a single partition of the data into clusters
- Good for large data sets
- Determining the number of clusters is a major challenge

K-means

Works with numeric data only

1) Pick a number (K) of cluster centers (at random)
2) Assign every item to its nearest cluster center (e.g. using Euclidean distance)
3) Move each cluster center to the mean of its assigned items
4) Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

K-Means

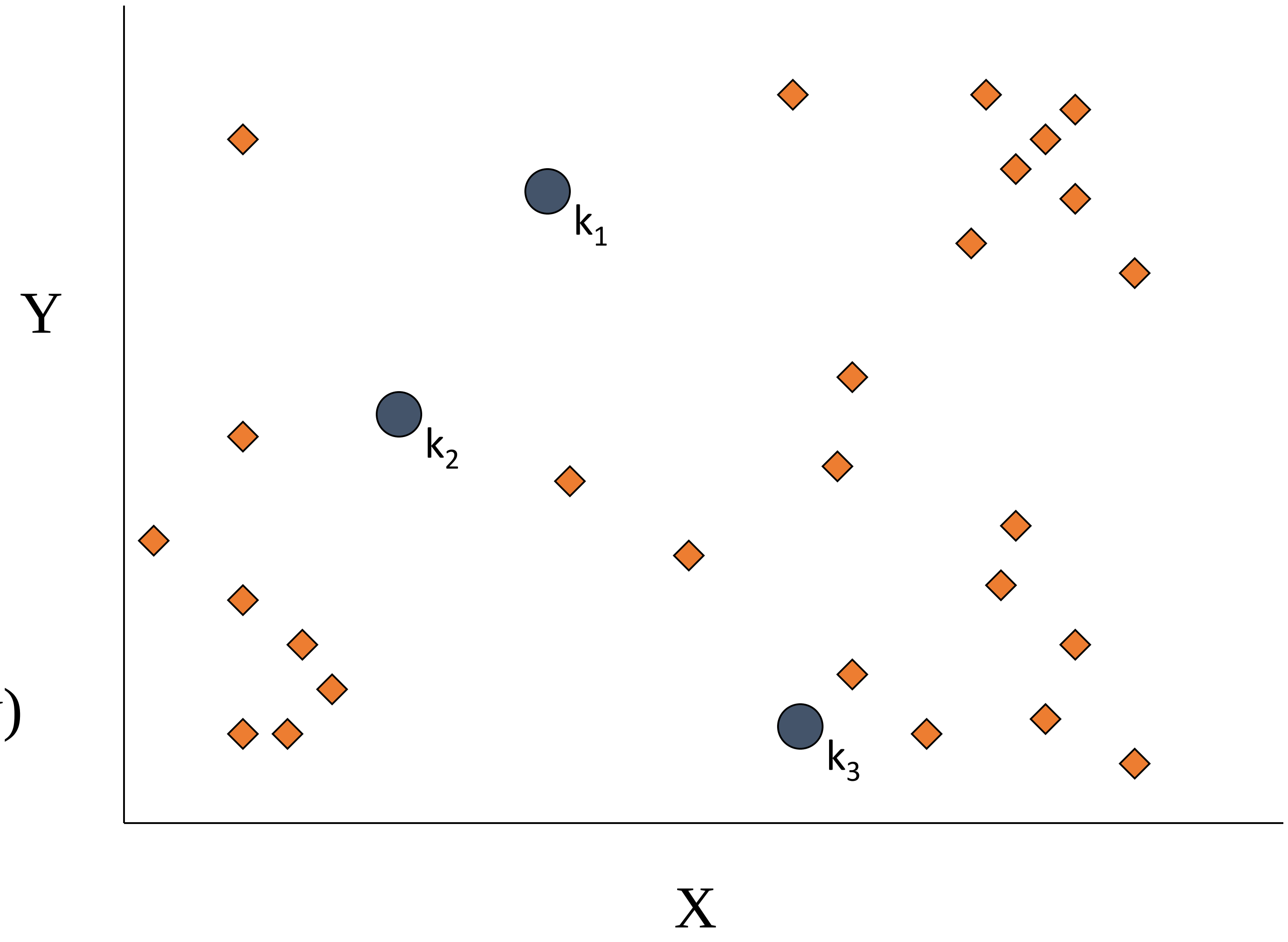
- Clusters based on *centroids* (aka the *center of gravity* or mean) of points in a cluster, *c*:

$$\vec{\mu}(\mathrm{c}) = \frac{1}{|c|}\sum_{\vec{x} \in c} \vec{x}$$

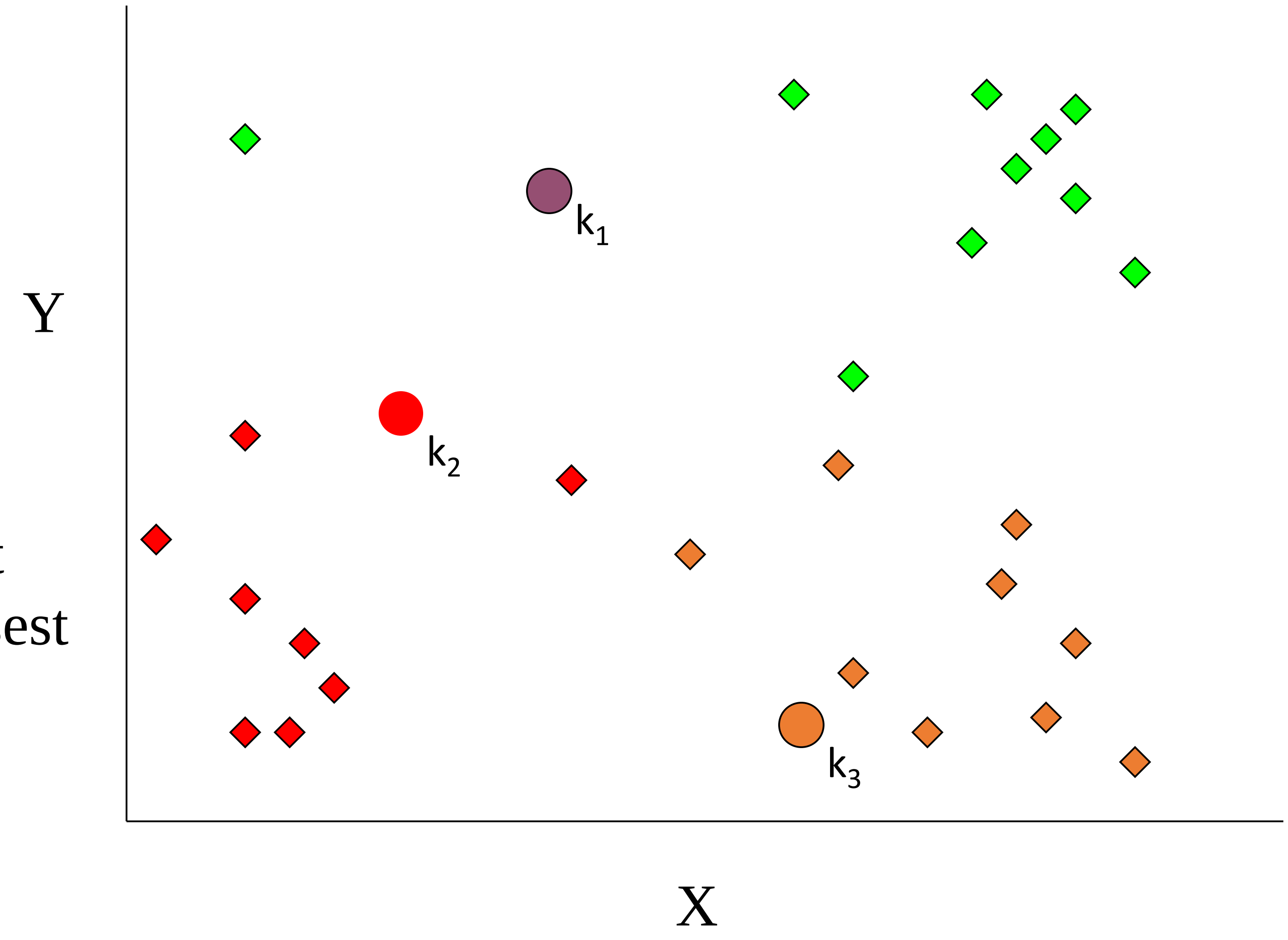
- Reassignment of instances to clusters is based on distance to the current cluster centroids.

K-means example, step 1

Pick 3 initial cluster centers (randomly)

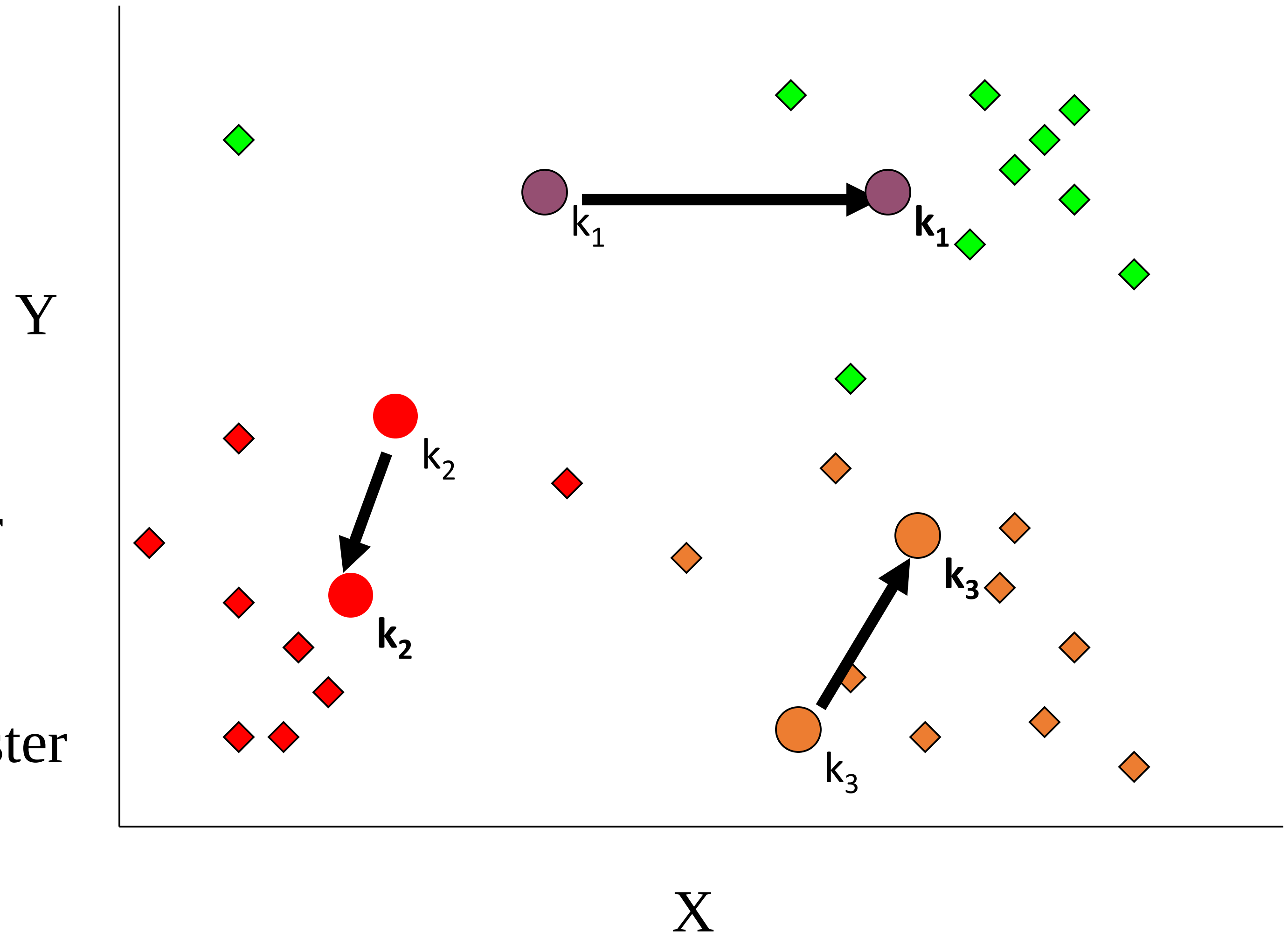


K-means example, step 2



Assign each point to the closest cluster center

K-means example, step 3

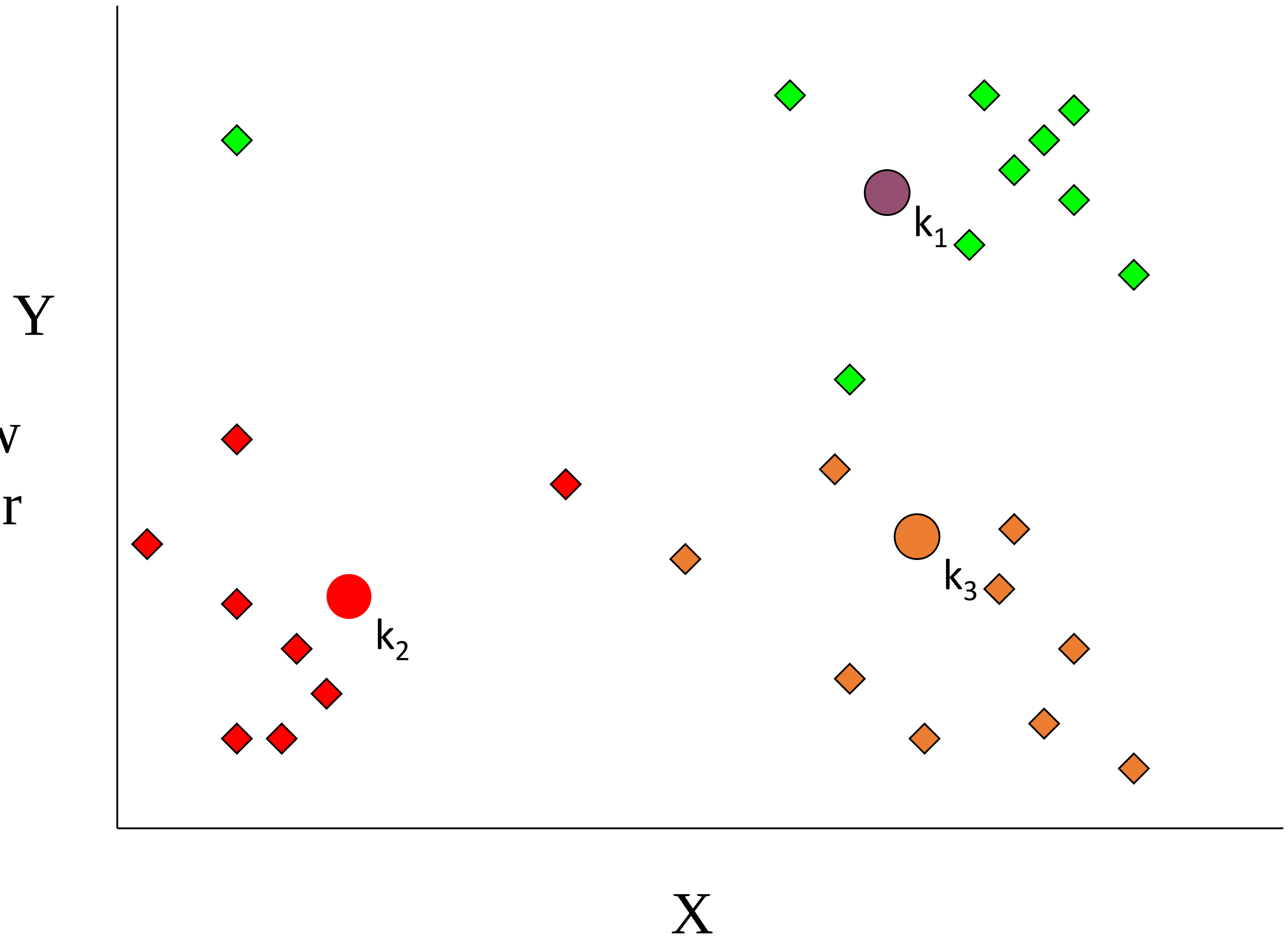


Move
each cluster
center
to the mean
of each cluster

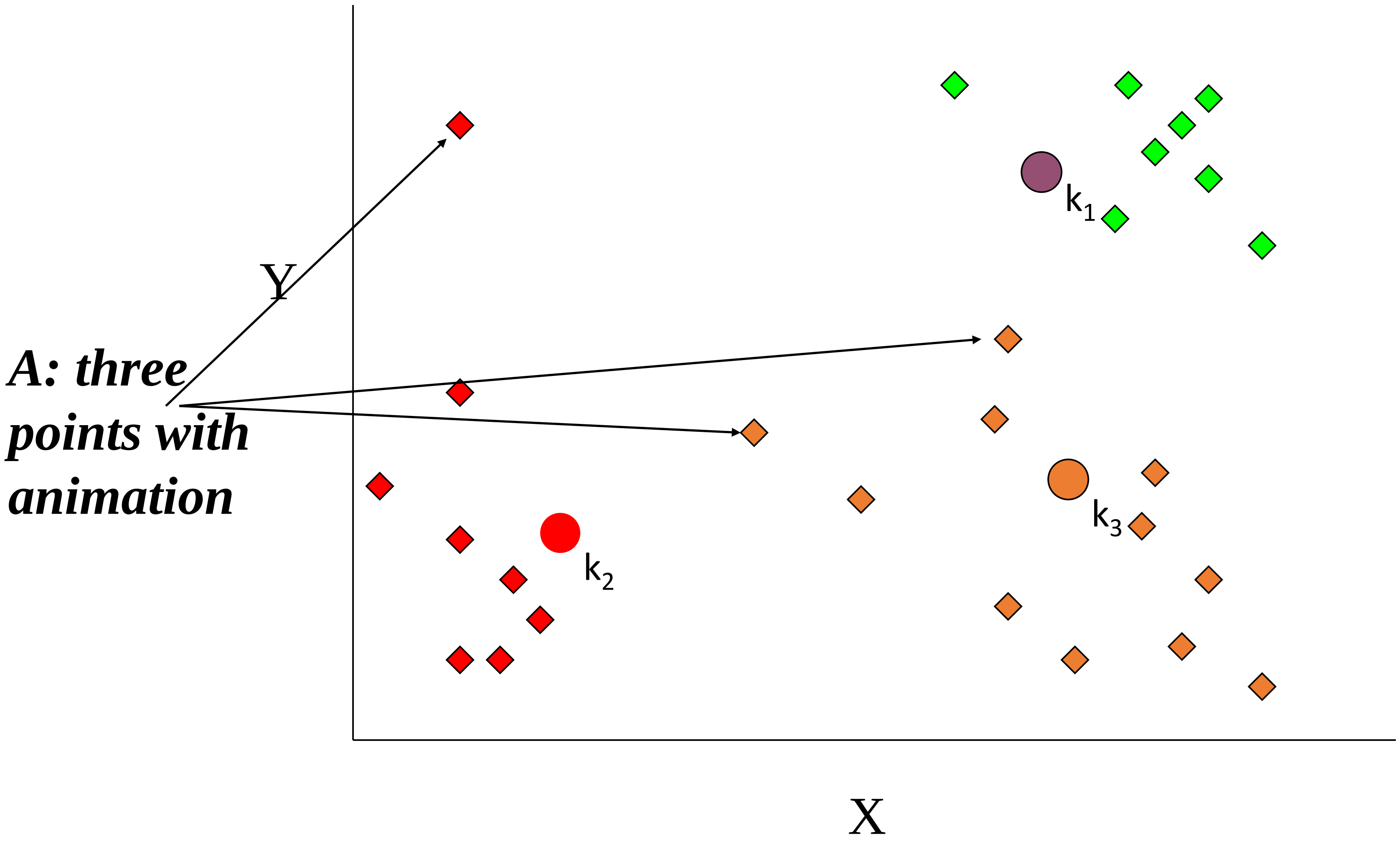
K-means example, step 4

Reassign points closest to a different new cluster center

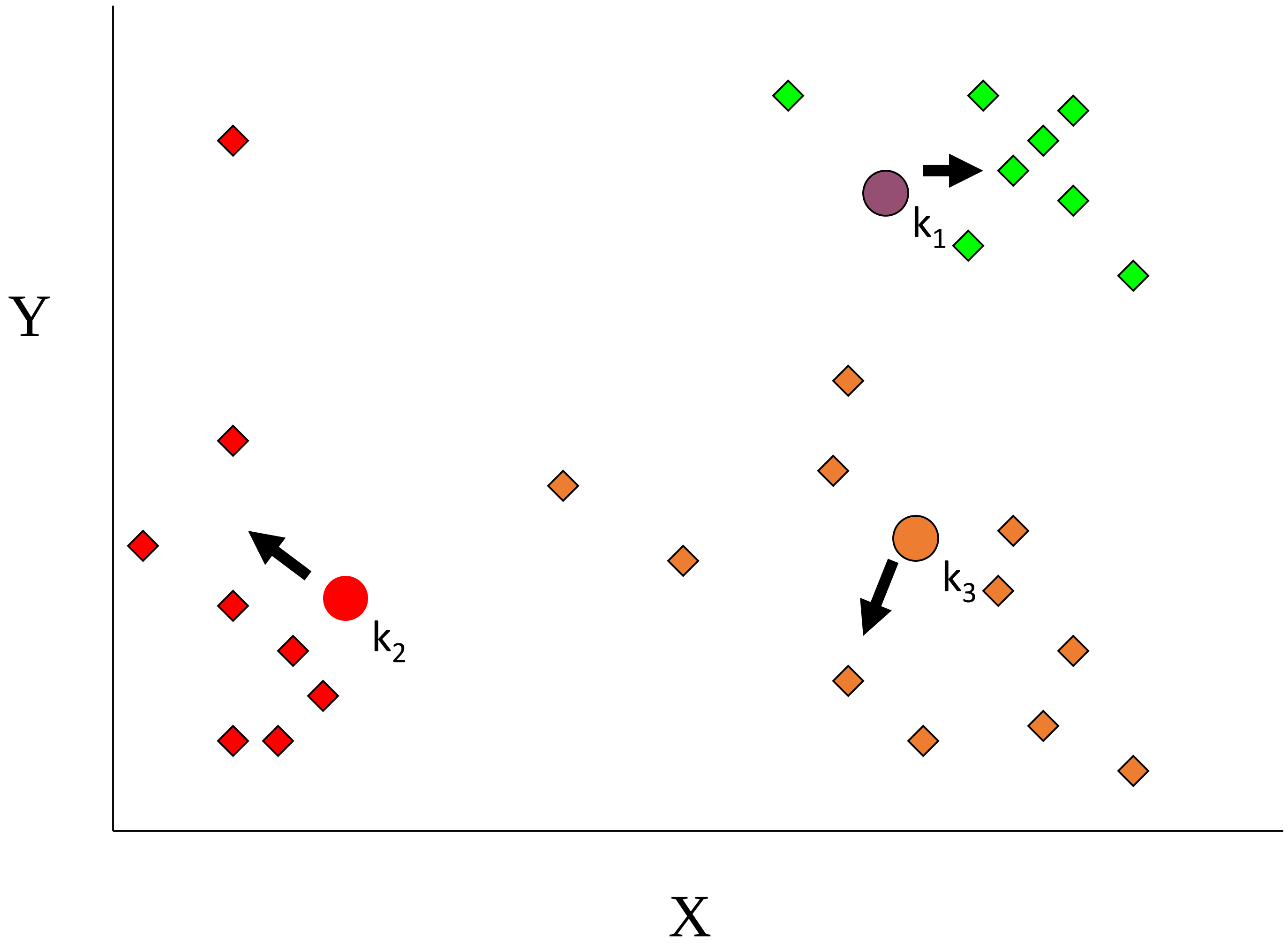
Q: Which points are reassigned?



K-means example, step 4 …

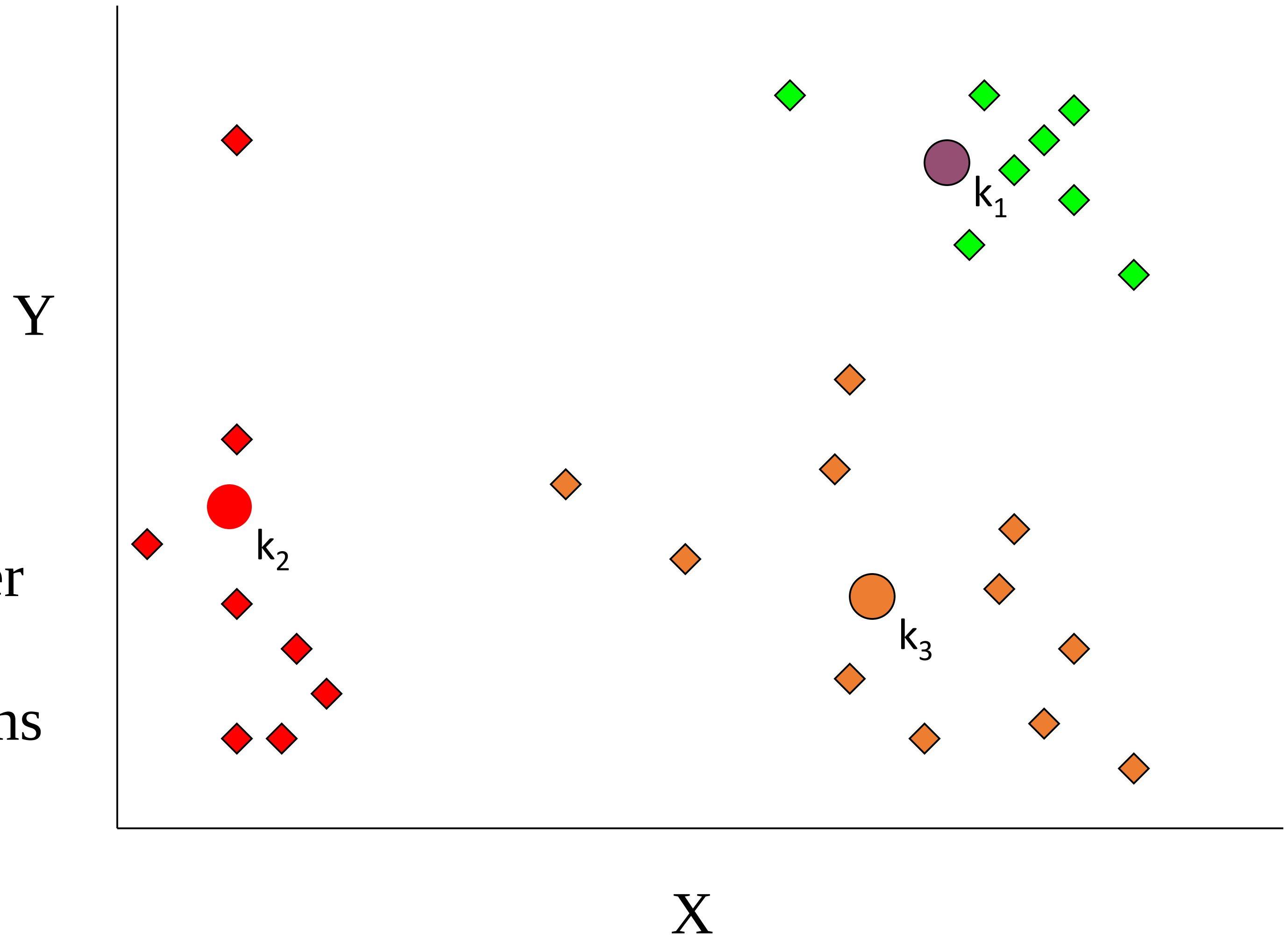


K-means example, step 4b



re-compute cluster means

K-means example, step 5



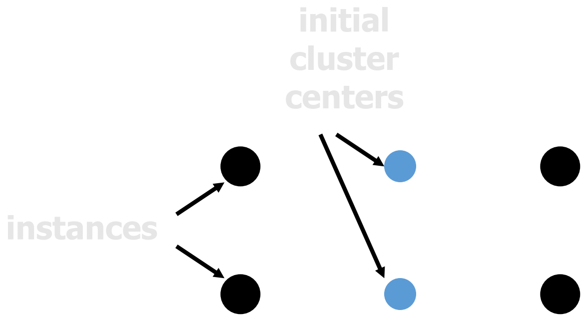
move cluster centers to cluster means

Discussion, 1

What can be the problems with
K-means clustering?

Discussion, 2

- Result can vary significantly depending on initial choice of seeds (number and position)
- Can get trapped in local minimum
  - Example:



- Q: What can be done?

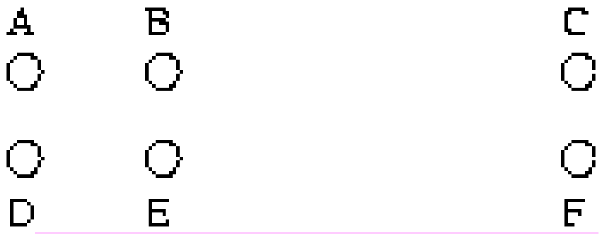
Discussion, 3

A: To increase chance of finding global optimum: restart with different random seeds.

Seed Choice

- Results can vary based on random seed selection.
- Some seeds can result in poor convergence rate, or convergence to sub-optimal clusterings.
  - Select good seeds using a heuristic (e.g., doc least similar to any existing mean)
  - Try out multiple starting points
  - Initialize with the results of another method.

Example showing sensitivity to seeds



In the above, if you start with B and E as centroids you converge to {A,B,C} and {D,E,F}
If you start with D and F you converge to {A,B,D,E} {C,F}

K-means clustering - outliers ?

What can be done about outliers?

K-means variations

- **K-medoids** – instead of mean, use medians of each cluster
  - Mean of 1, 3, 5, 7, 9 is 5
  - Mean of 1, 3, 5, 7, 1009 is 205
  - Median of 1, 3, 5, 7, 1009 is 5
  - Median advantage: not affected by extreme values
- For large databases, use sampling

How Many Clusters?

- Number of clusters K is given
  - Partition n docs into predetermined number of clusters
- Finding the “right” number of clusters is part of the problem
  - Given data, partition into an “appropriate” number of subsets.
  - E.g., for query results - ideal value of K not known up front - though UI may impose limits.
- Can usually take an algorithm for one flavor and convert to the other.

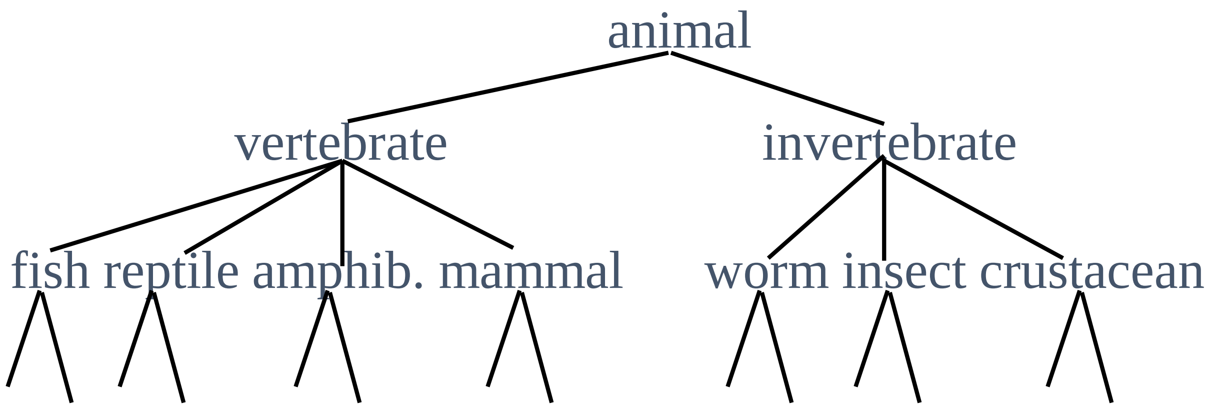
How Many Clusters?

- http://www.ee.columbia.edu/~dpwe/papers/PhamDN05-kmeans.pdf
- http://web.stanford.edu/~hastie/Papers/gap.pdf

Hierarchical clustering

Hierarchical Clustering

- Build a tree-based hierarchical taxonomy (*dendrogram*) from a set of documents.



How could you do this with k-means?

Hierarchical Clustering algorithms

- **Agglomerative (bottom-up):**
  - Start with each document being a single cluster.
  - Eventually all documents belong to the same cluster.
- **Divisive (top-down):**
  - Start with all documents belong to the same cluster.
  - Eventually each node forms a cluster on its own.
  - Could be a recursive application of k-means like algorithms
- Does not require the number of clusters ***k*** in advance
- Needs a termination/readout condition

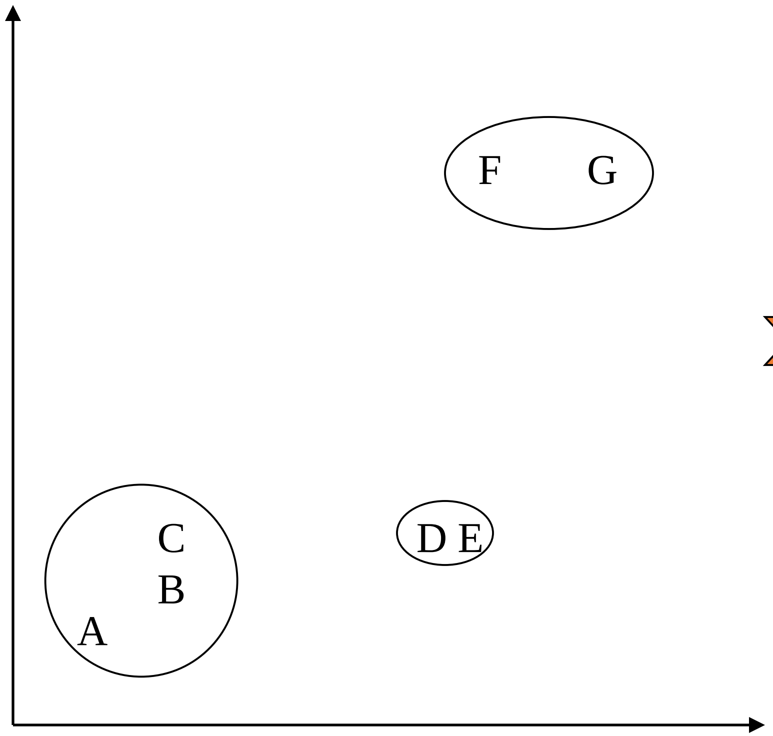
Hierarchical Clustering algorithms

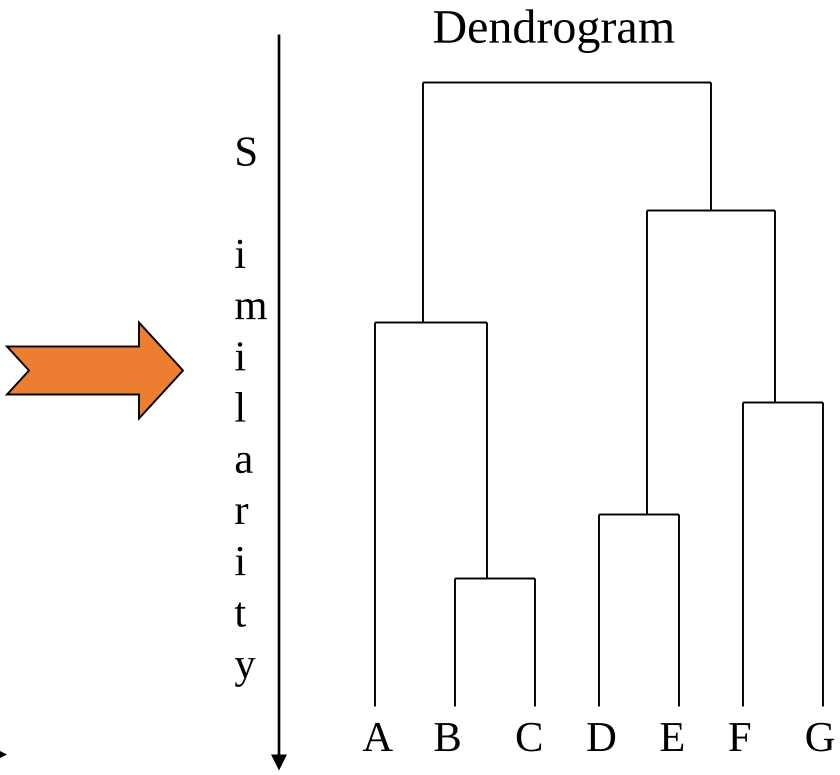
- **Agglomerative (bottom-up):**
  - Start with each document being a single cluster.
  - Eventually all documents belong to the same cluster.
- **Divisive (top-down):**
  - Start with all documents belong to the same cluster.
  - Eventually each node forms a cluster on its own.
  - Could be a recursive application of k-means like algorithms
- Does not require the number of clusters ***k*** in advance
- Needs a termination/readout condition

Hierarchical Agglomerative Clustering (HAC)

- Assumes a similarity function for determining the similarity of two instances.
- Starts with all instances in a separate cluster and then repeatedly joins the two clusters that are most similar until there is only one cluster.
- The history of merging forms a binary tree or hierarchy.

Hierarchical Clustering

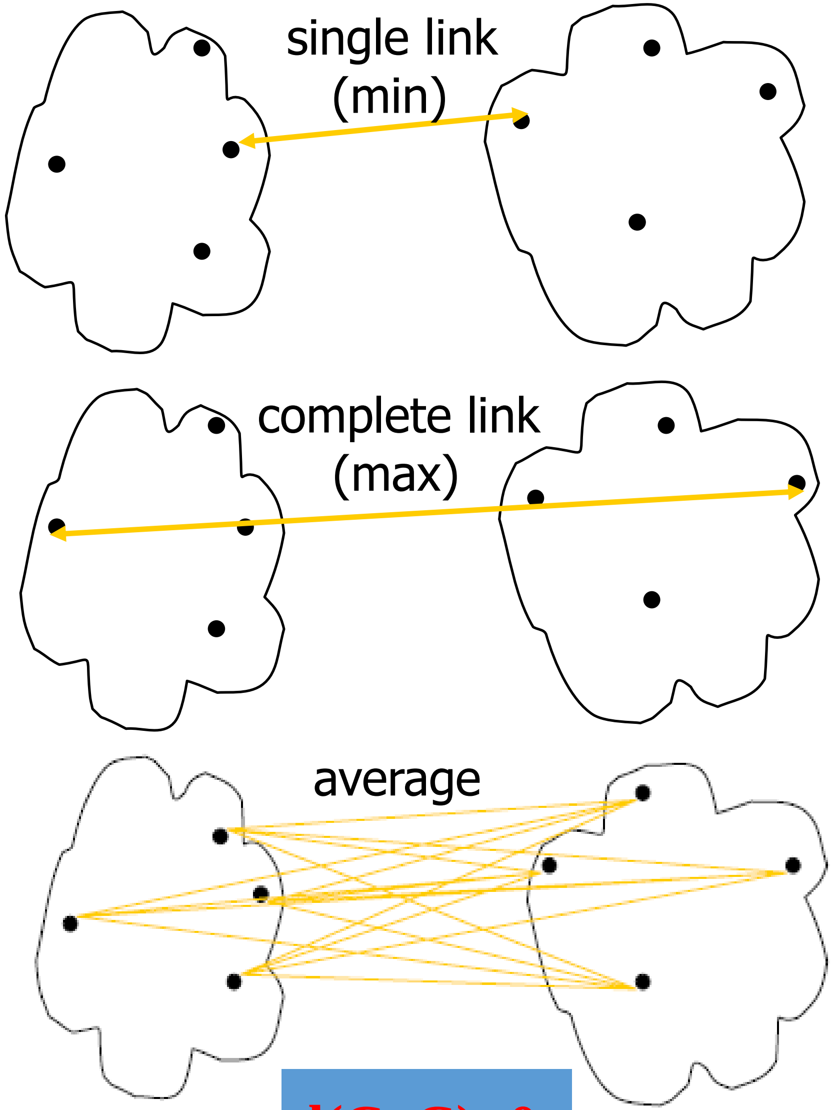




Cluster Distance Measures

- **Single link**: smallest distance between an element in one cluster and an element in the other, i.e., $d(C_i, C_j) = \min\{d(x_{ip}, x_{jq})\}$
- **Complete link**: largest distance between an element in one cluster and an element in the other, i.e., $d(C_i, C_j) = \max\{d(x_{ip}, x_{jq})\}$
- **Average**: avg distance between elements in one cluster and elements in the other, i.e.,

$d(C_i, C_j) = \text{avg}\{d(x_{ip}, x_{jq})\}$



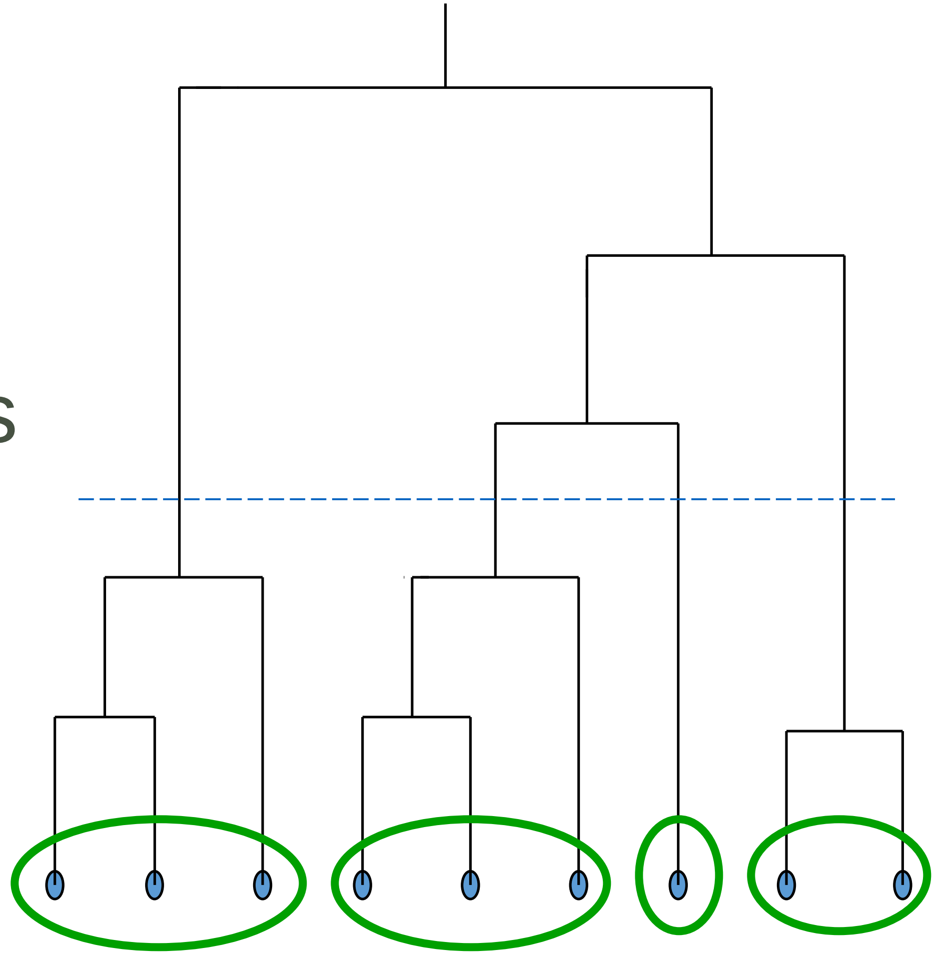
d(C, C)=0

Ward’s hierarchical clustering

- Goal: minimize ”Error Sum of Squares” (ESS) at every step.
  - ESS = The sum over all clusters, of the sum of the squares of the distances from the objects to the cluster centroid.
- When joining two clusters, find the pair that results in the smallest increase in ESS.

Dendogram: Hierarchical Clustering

- Clustering obtained by cutting the dendrogram at a desired level: each connected component forms a cluster.



Hierarchical Agglomerative Clustering (HAC)

- Starts with each doc in a separate cluster
  - then repeatedly joins the *closest pair* of clusters, until there is only one cluster.
- The history of merging forms a binary tree or hierarchy.

How to measure distance of clusters??

Closest pair of clusters

Many variants to defining closest pair of clusters

- **Single-link**
  - Distance of the *"closest" points* (single-link)
- **Complete-link**
  - Distance of the "furthest" points
- **Centroid**
  - Distance of the centroids (centers of gravity)
- **(Average-link)**
  - Average distance between pairs of elements

Single Link Agglomerative Clustering

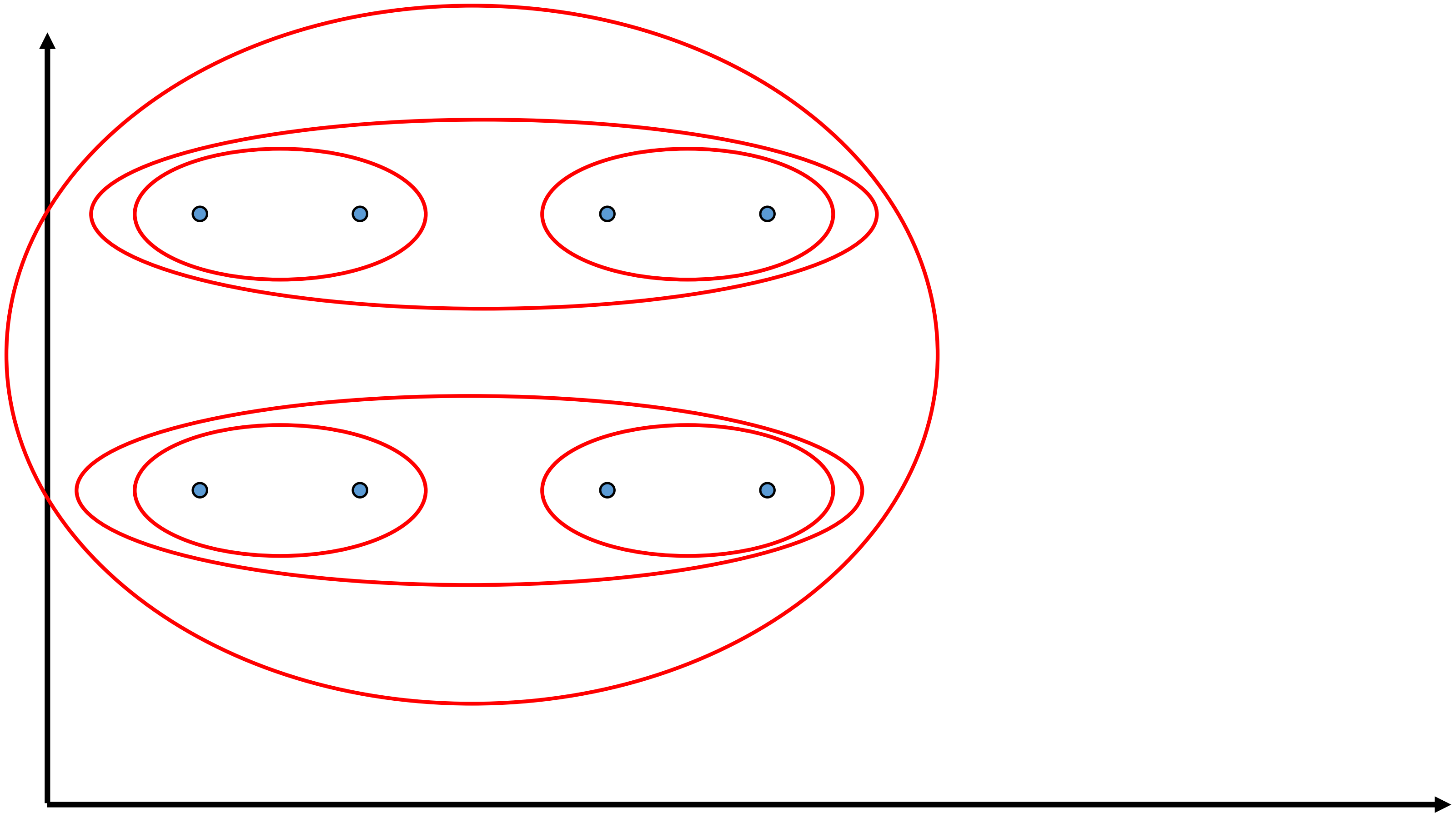
- Use maximum similarity of pairs:

$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$

- Can result in “straggly” (long and thin) clusters due to chaining effect.
- After merging c_i and c_j, the similarity of the resulting cluster to another cluster, c_k, is:

$$sim((c_i \cup c_j), c_k) = \max(sim(c_i, c_k), sim(c_j, c_k))$$

Single Link Example



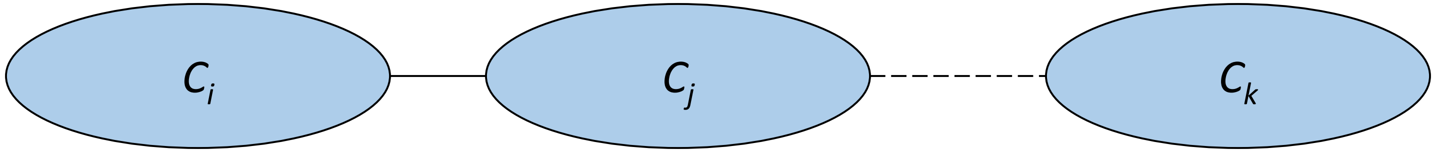
Complete Link Agglomerative Clustering

- Use minimum similarity of pairs:

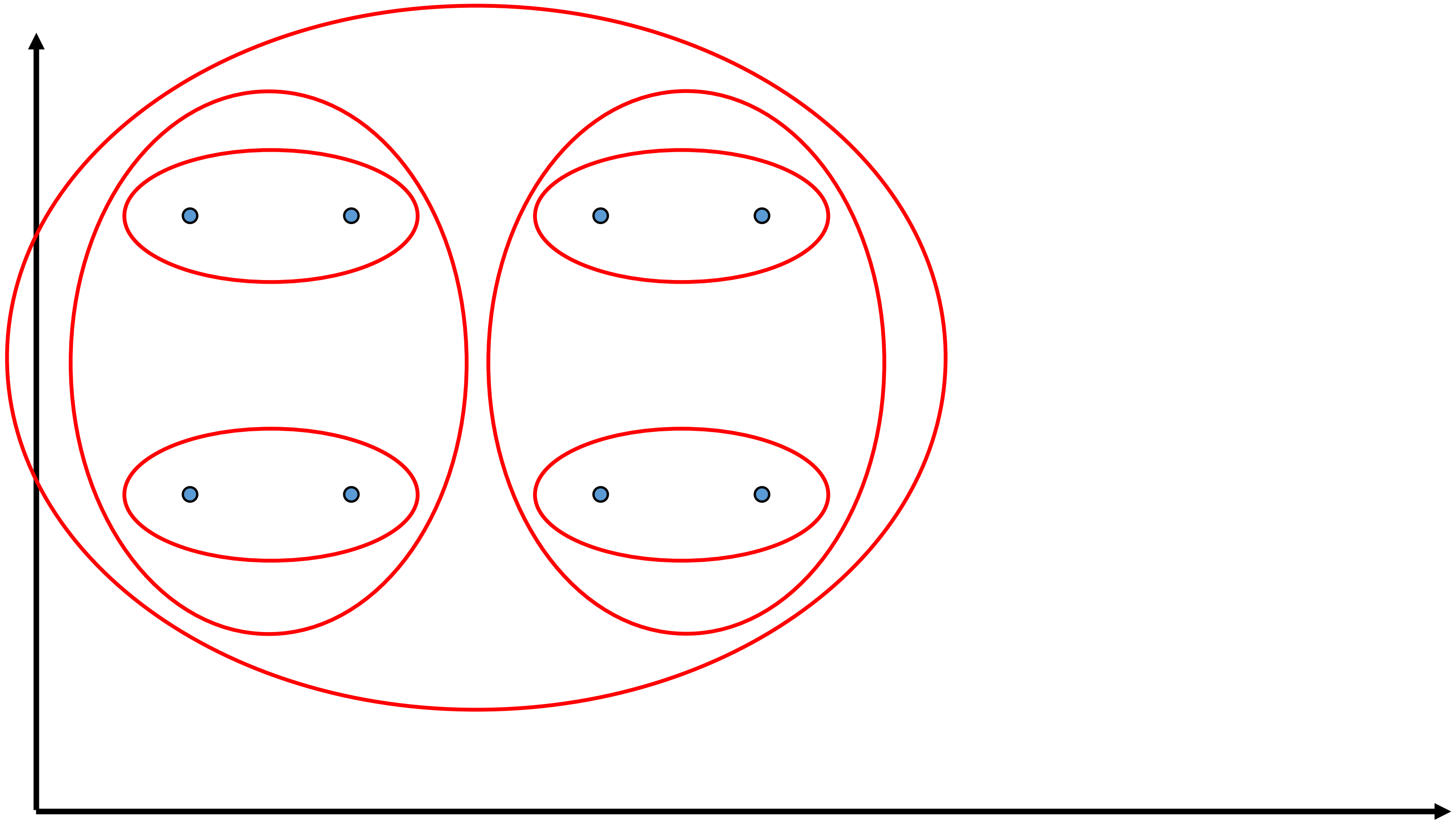
$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$

- Makes "tighter," spherical clusters that are typically preferable.
- After merging c_i and c_j, the similarity of the resulting cluster to another cluster, c_k, is:

$$sim((c_i \cup c_j), c_k) = \min(sim(c_i, c_k), sim(c_j, c_k))$$



Complete Link Example



Key notion: *cluster representative*

- We want a notion of a representative point in a cluster
- Representative should be some sort of “typical” or central point in the cluster, e.g.,
  - point inducing smallest radii to docs in cluster
  - smallest squared distances, etc.
  - point that is the “average” of all docs in the cluster
    - Centroid or center of gravity

Centroid-based Similarity

- Always maintain average of vectors in each cluster:

$$\vec{s}(c_j) = \frac{\sum_{\vec{x} \in c_j} \vec{x}}{\left|c_j\right|}$$

- Compute similarity of clusters by:

$$sim(c_i, c_j) = sim(s(c_i), s(c_j))$$

- For non-vector data, can’t always make a centroid

Computational Complexity

- In the first iteration, all HAC methods need to compute similarity of all pairs of *n* individual.
- In each of the subsequent n–2 merging iterations, compute the distance between the most recently created cluster and all other existing clusters.

Major issue - labeling

- After clustering algorithm finds clusters - how can they be useful to the end user?
- Need pithy label for each cluster

Other Clustering

- Artificial Neural Networks (ANN)
- Random search
  - Genetic Algorithms (GA)
    - GA used to find initial centroids for *k*-means
  - Simulated Annealing (SA)
  - Tabu Search (TS)
- Support Vector Machines (SVM)
- Fuzzy

Probabilistic or fuzzy clustering

- The output is, for each object and each cluster, a probability or weight that the object belongs to the cluster
- Example: The observations are modelled as produced by drawing from a number of probability densities (often multivariate normal). Parameters are then estimated with Maximum Likelihood (for example using EM algorithm).
- Example: A ”fuzzy” version of k-means, where weights for objects are changed iteratively

Neural networks for clustering

- Neural networks are mathematical models made to be similar to actual neural networks
- They consist of layers of nodes that send out ”signals” based probabilistically on input signals
- Most known uses are classifications, i.e., with learning sets

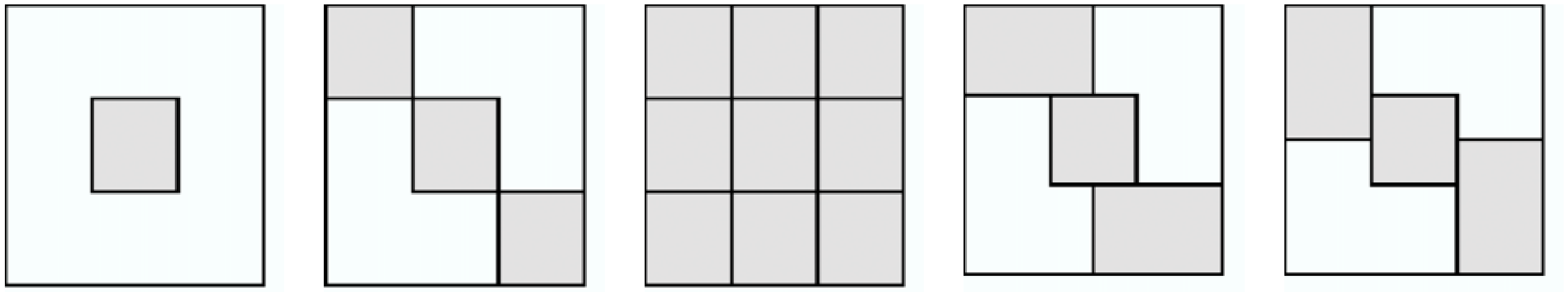
Biclustering

- A biclustering method is an unsupervised learning method which looks for sub-matrices in a data matrix with a high similarity of elements.
- Algorithms: Statistical based, AI, machine learning.
- BiclustGUI: A User Friendly Interface for Biclustering Analysis

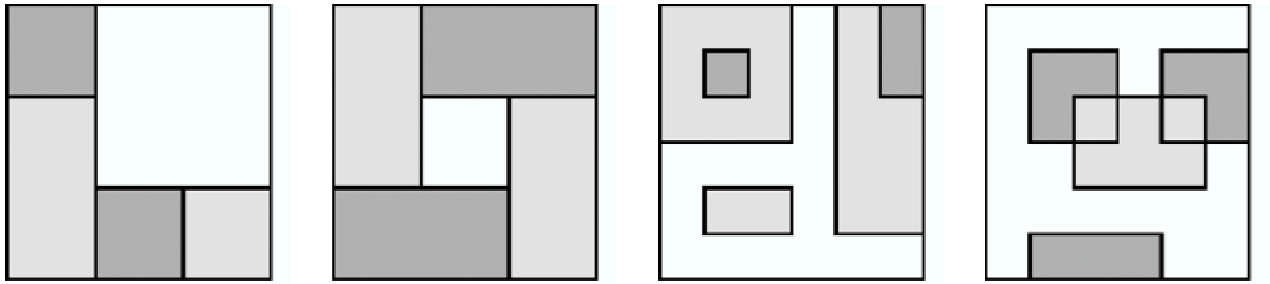
Biclustering in Bioinformatics

- Genes not regulated under all conditions
- Genes regulated by multiple factors/processes concurrently
- Key to determine function of genes
- Key to determine classification of conditions

Bicluster Structure



(a) Single Bicluster (b) Exclusive row and column biclusters (c) Checkerboard Structure (d) Exclusive-rows biclusters (e) Exclusive-columns biclusters



(f) Non-Overlapping biclusters with tree structure (g) Non-Overlapping non-exclusive biclusters (h) Overlapping biclusters with hierarchical structure (i) Arbitrarily positioned overlapping biclusters

What is a Good Clustering?

What is a Good Clustering?

- *Internal criterion*: A good clustering will produce high quality clusters in which:
  - the intra-class (that is, intra-cluster) similarity is high
  - the inter-class similarity is low
  - The measured quality of a clustering depends on both the document representation and the similarity measure used

External criteria for clustering quality

- Quality measured by its ability to discover some or all of the hidden patterns or latent classes in gold standard data
- Assesses a clustering with respect to ground truth
- Assume documents with C gold standard classes, while our clustering algorithms produce K clusters, $\omega_1, \omega_2, \ldots, \omega_K$ with n_i members.

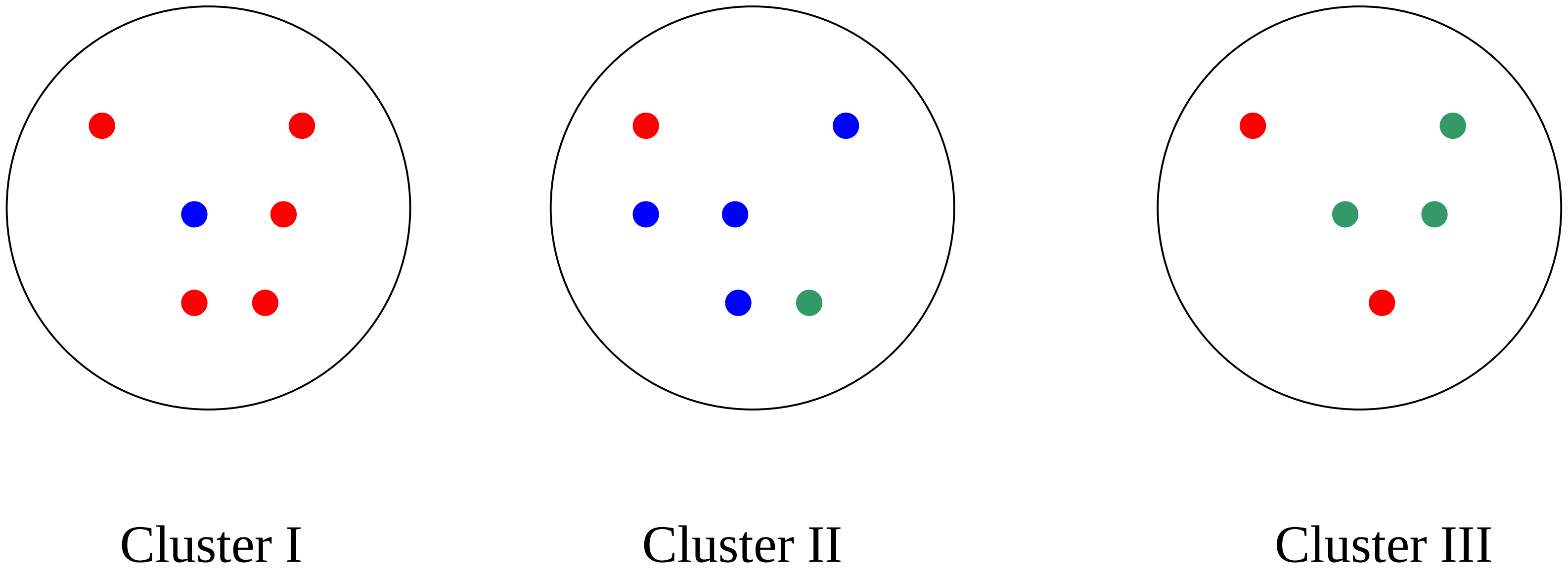
External Evaluation of Cluster Quality

- *Simple measure*: purity, the ratio between the dominant class in the cluster π_i and the size of cluster ω_i

$$Purity(\omega_i) = \frac{1}{n_i} \max_j (n_{ij}) \quad j \in C$$

- Others are entropy of classes in clusters (or mutual information between classes and clusters)

Purity example



Cluster I: Purity = 1/6 (max(5, 1, 0)) = 5/6

Cluster II: Purity = 1/6 (max(1, 4, 1)) = 4/6

Cluster III: Purity = 1/5 (max(2, 0, 3)) = 3/5

Rand Index

Number of points	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	A	C
Different classes in ground truth	B	D

Rand index: symmetric version

$$RI = \frac{A+D}{A+B+C+D}$$

Compare with standard Precision and Recall.

$$P = \frac{A}{A+B} \qquad R = \frac{A}{A+C}$$

Rand Index example: 0.68

Number of points	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	20	24
Different classes in ground truth	20	72

Evaluation of clustering

- Perhaps the most substantive issue in data mining in general:
  - how do you measure goodness?
- Most measures focus on computational efficiency
  - Time and space
- For application of clustering to search:
  - Measure retrieval effectiveness

Approaches to evaluating

- Anecdotal
- User inspection
- Ground “truth” comparison
  - Cluster retrieval
- Purely quantitative measures
  - Probability of generating clusters found
  - Average distance between cluster members
- Microeconomic / utility

Anecdotal evaluation

- Probably the commonest (and surely the easiest)
  - “I wrote this clustering algorithm and look what it found!”
- No benchmarks, no comparison possible
- Any clustering algorithm will pick up the easy stuff like partition by languages
- Generally, unclear scientific value.

User inspection

- Induce a set of clusters or a navigation tree
- Have subject matter experts evaluate the results and score them
  - some degree of subjectivity
- Often combined with search results clustering
- Not clear how reproducible across tests.
- Expensive / time-consuming

Thank you….