Spring 2019

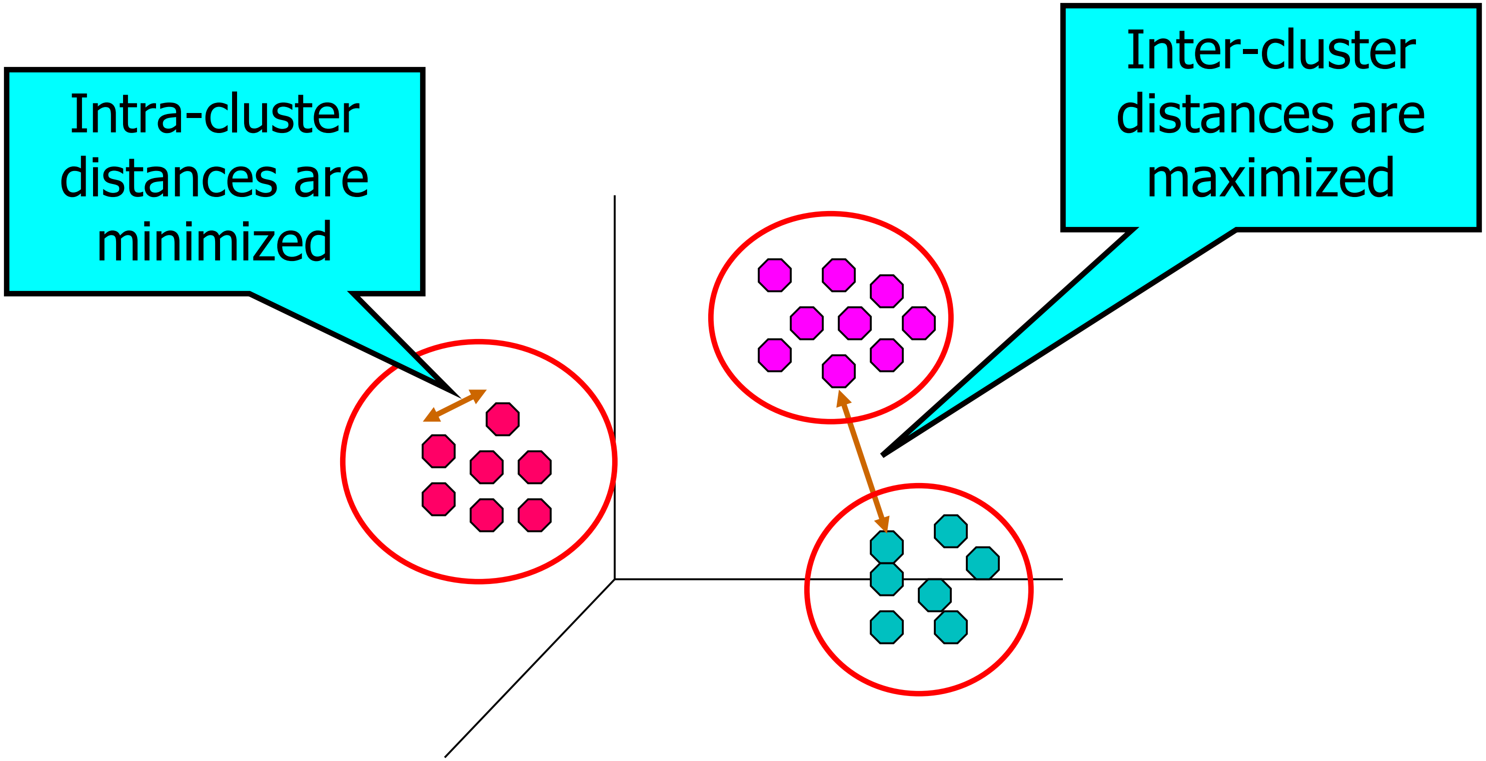
CSEN1083: Data Mining

Clustering Analysis

Seif Eldawlatly

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

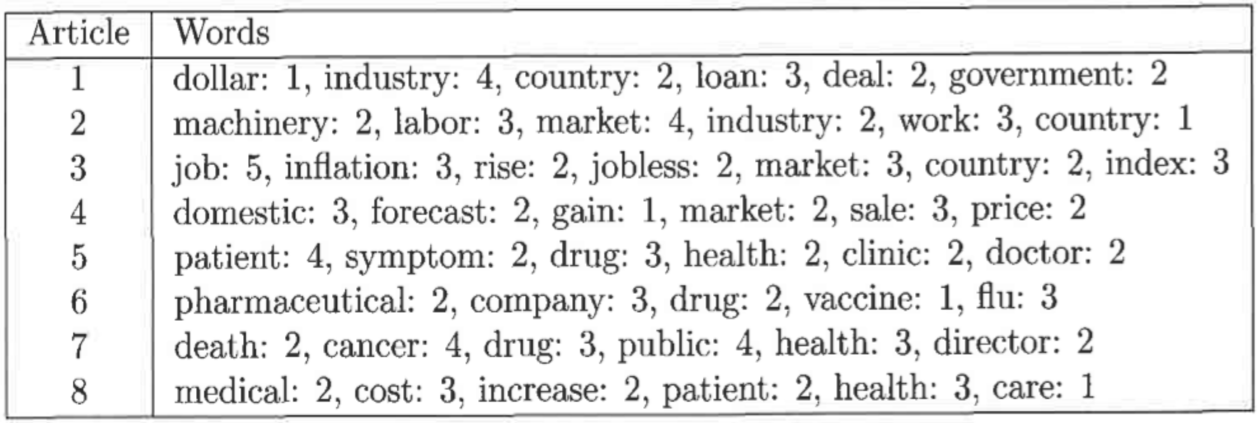


Intra-cluster distances are minimized

Inter-cluster distances are maximized

Clustering

- Example: Understanding
  - Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations

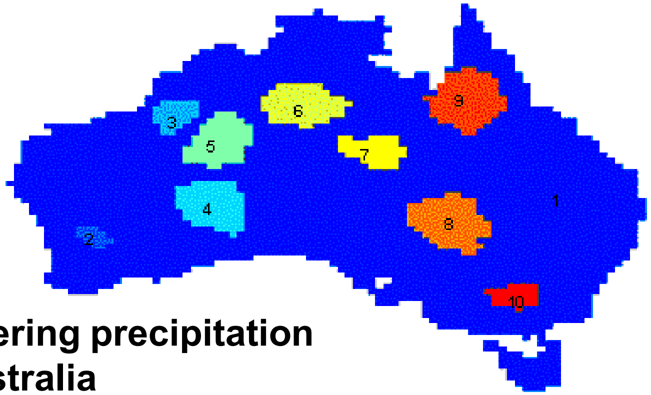


Article	Words
1	dollar: 1, industry: 4, country: 2, loan: 3, deal: 2, government: 2
2	machinery: 2, labor: 3, market: 4, industry: 2, work: 3, country: 1
3	job: 5, inflation: 3, rise: 2, jobless: 2, market: 3, country: 2, index: 3
4	domestic: 3, forecast: 2, gain: 1, market: 2, sale: 3, price: 2
5	patient: 4, symptom: 2, drug: 3, health: 2, clinic: 2, doctor: 2
6	pharmaceutical: 2, company: 3, drug: 2, vaccine: 1, flu: 3
7	death: 2, cancer: 4, drug: 3, public: 4, health: 3, director: 2
8	medical: 2, cost: 3, increase: 2, patient: 2, health: 3, care: 1

Cluster 1: Economy

Cluster 2: Healthcare

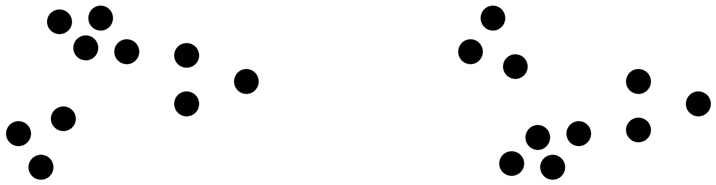
- Example: Summarization
  - Reduce the size of large data sets



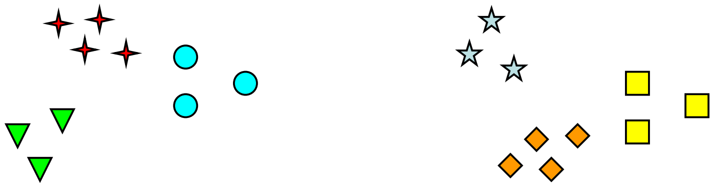
Clustering precipitation in Australia

Clustering

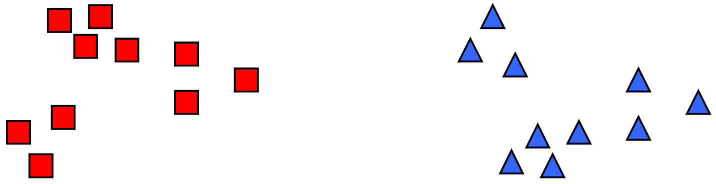
- The definition of a cluster is imprecise and the best definition depends on the nature of data and the desired results



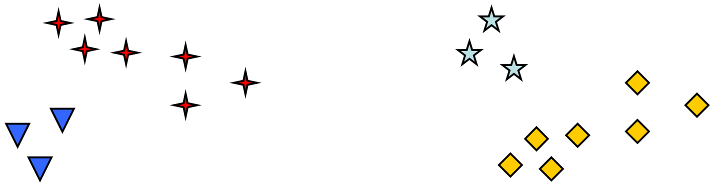
How many clusters?



Six Clusters



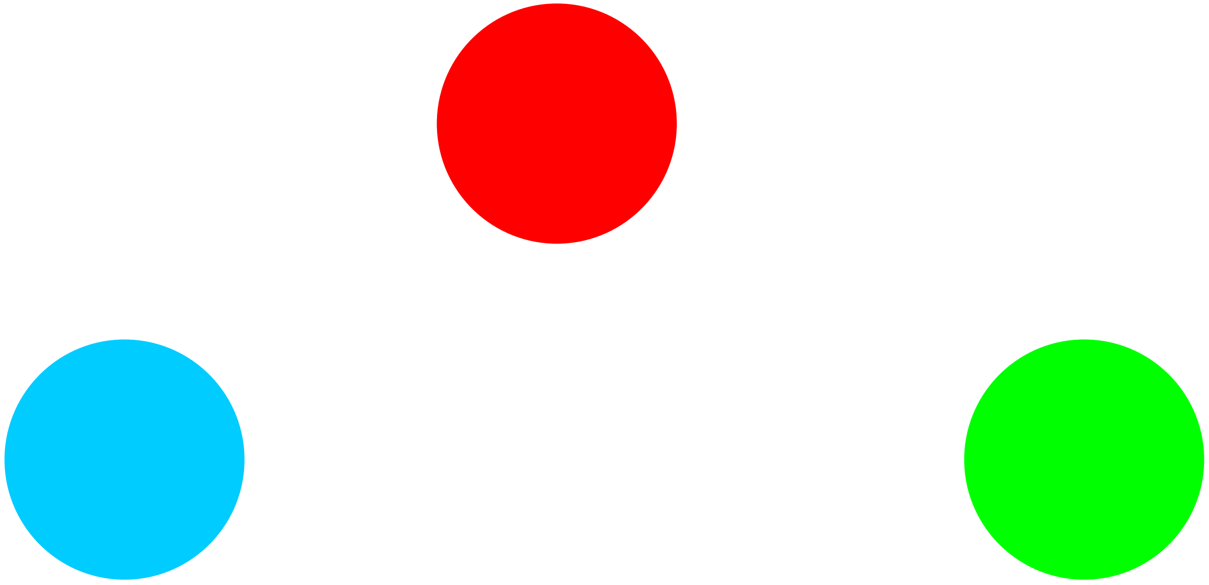
Two Clusters



Four Clusters

Types of Clusters

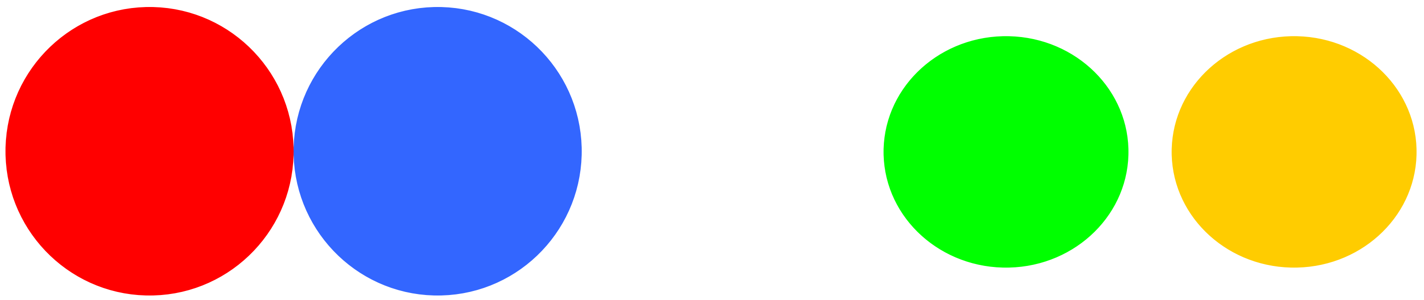
- Well-Separated Clusters:
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster



3 well-separated clusters

Types of Clusters

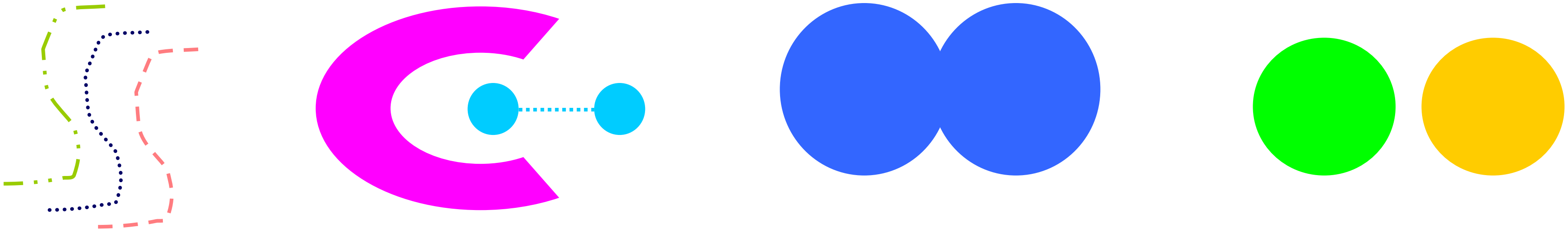
- Center-based
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster



4 center-based clusters

Types of Clusters

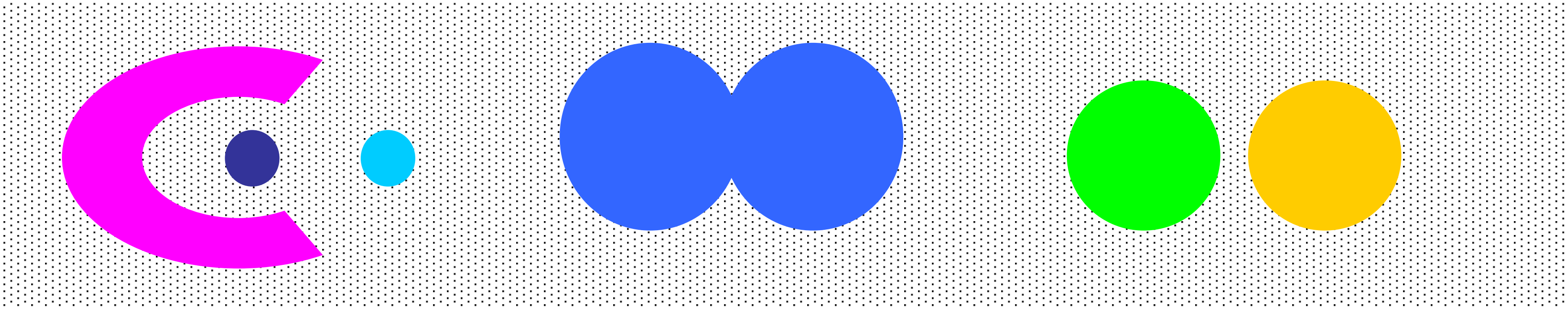
- Contiguous Cluster (Nearest neighbor or Transitive)
  - A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster



8 contiguous clusters

Types of Clusters

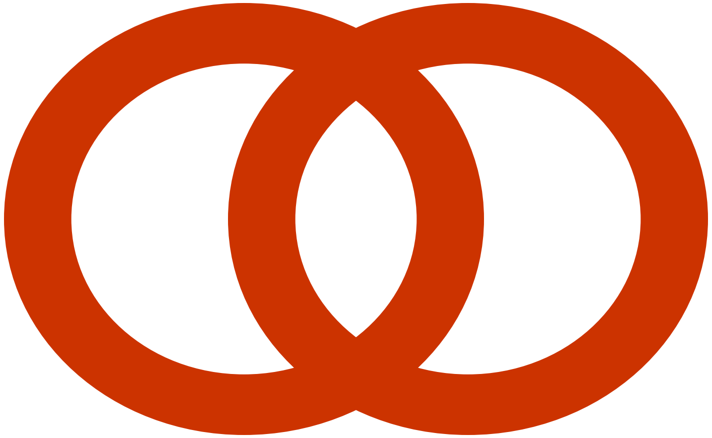
- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density



6 density-based clusters

Types of Clusters

- Shared Property or Conceptual Clusters
  - Finds clusters that share some common property or represent a particular concept.



2 Overlapping Circles

K-means Clustering

- Objective Function: Minimize J

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

$\mathbf{x}_n$: Input data

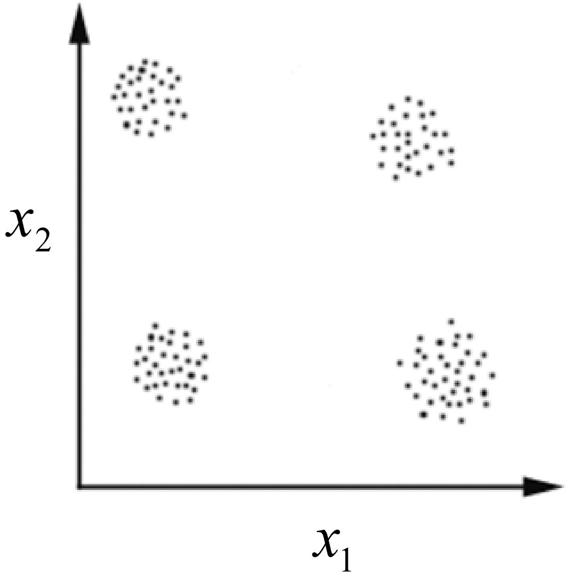
$\boldsymbol{\mu}_k$: Center of cluster k

r_{nk}: Cluster membership $= 1$ if $\mathbf{x}_n \in C_k$

$= 0$ if $\mathbf{x}_n \notin C_k$

N : Number of data points

K : Number of clusters to look for



K-means Clustering

- Algorithm steps:
  - Step 1: Randomly choose clusters center $\boldsymbol{\mu}_k$
  - Step 2: Compute r_{nk}

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

(Assign $\mathbf{x}_n$ to the cluster with closest center)

  - Step 3: Update $\boldsymbol{\mu}_k$

Take derivative of J w.r.t. $\boldsymbol{\mu}_k$ and equate with zero

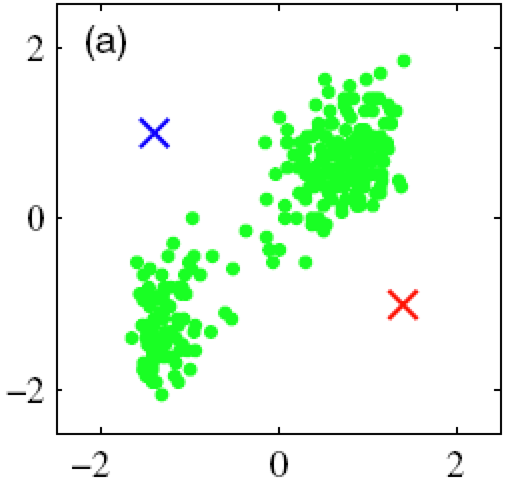
$$2\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \quad \rightarrow \quad \boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

  - Back to Step 2 until convergence

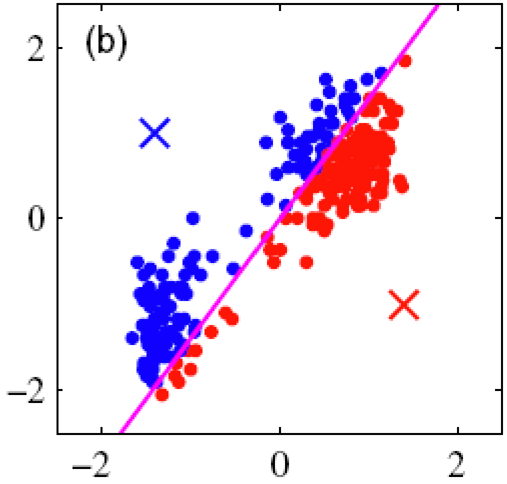
K-means Clustering

- Example

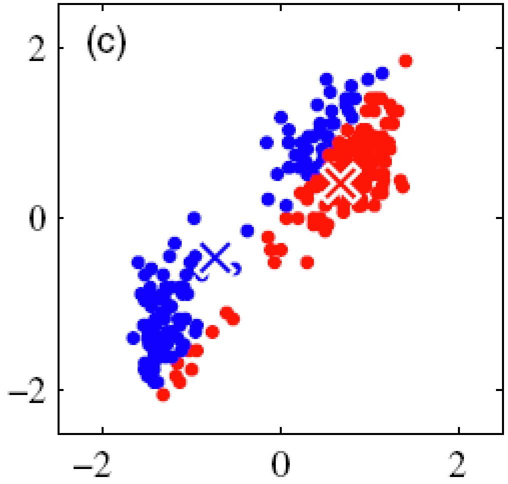
Randomly choose μ_k



Compute r_{nk}

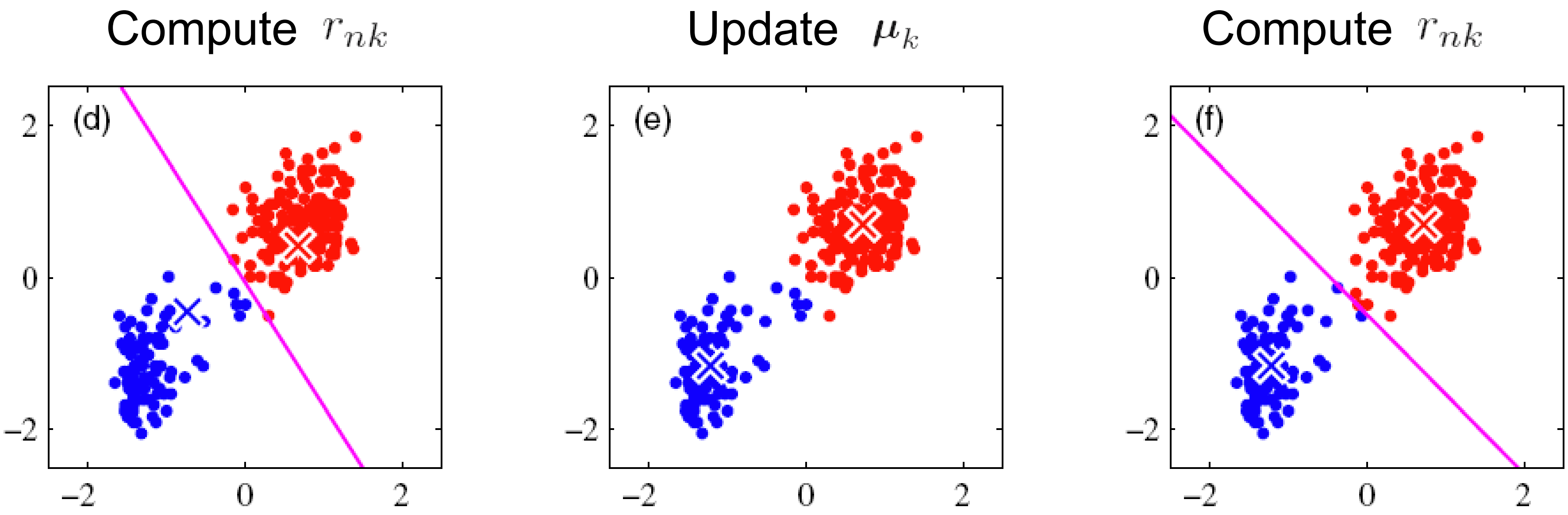


Update μ_k



K-means Clustering

- Example



Compute r_{nk}

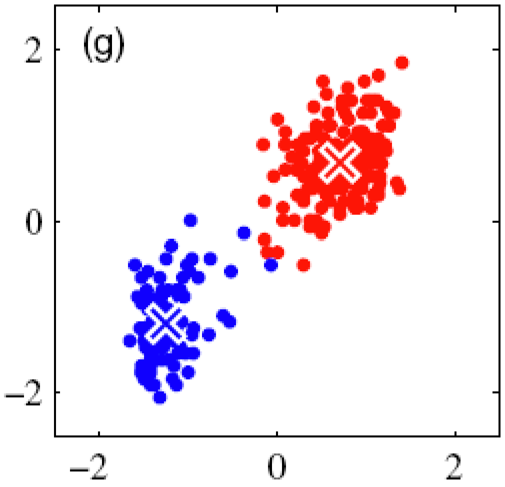
Update μ_k

Compute r_{nk}

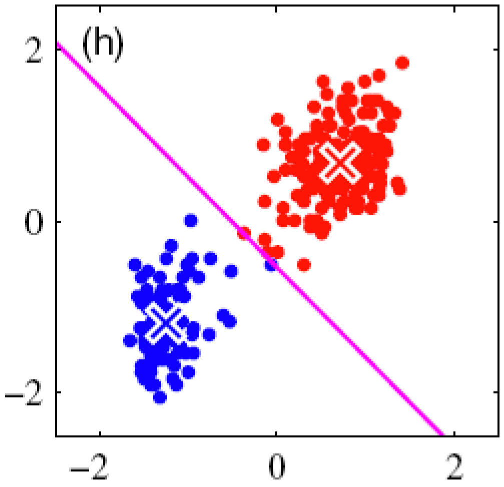
K-means Clustering

- Example

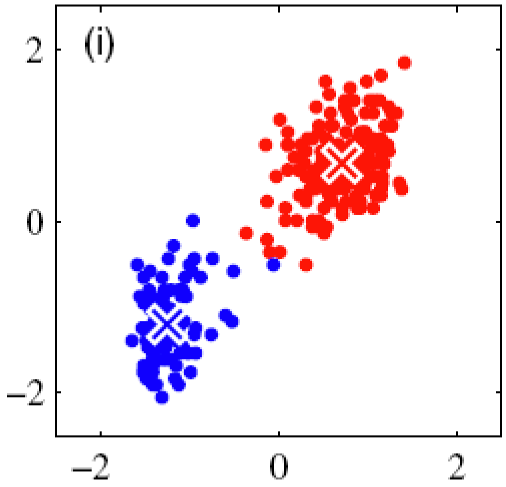
Update μ_k



Compute r_{nk}

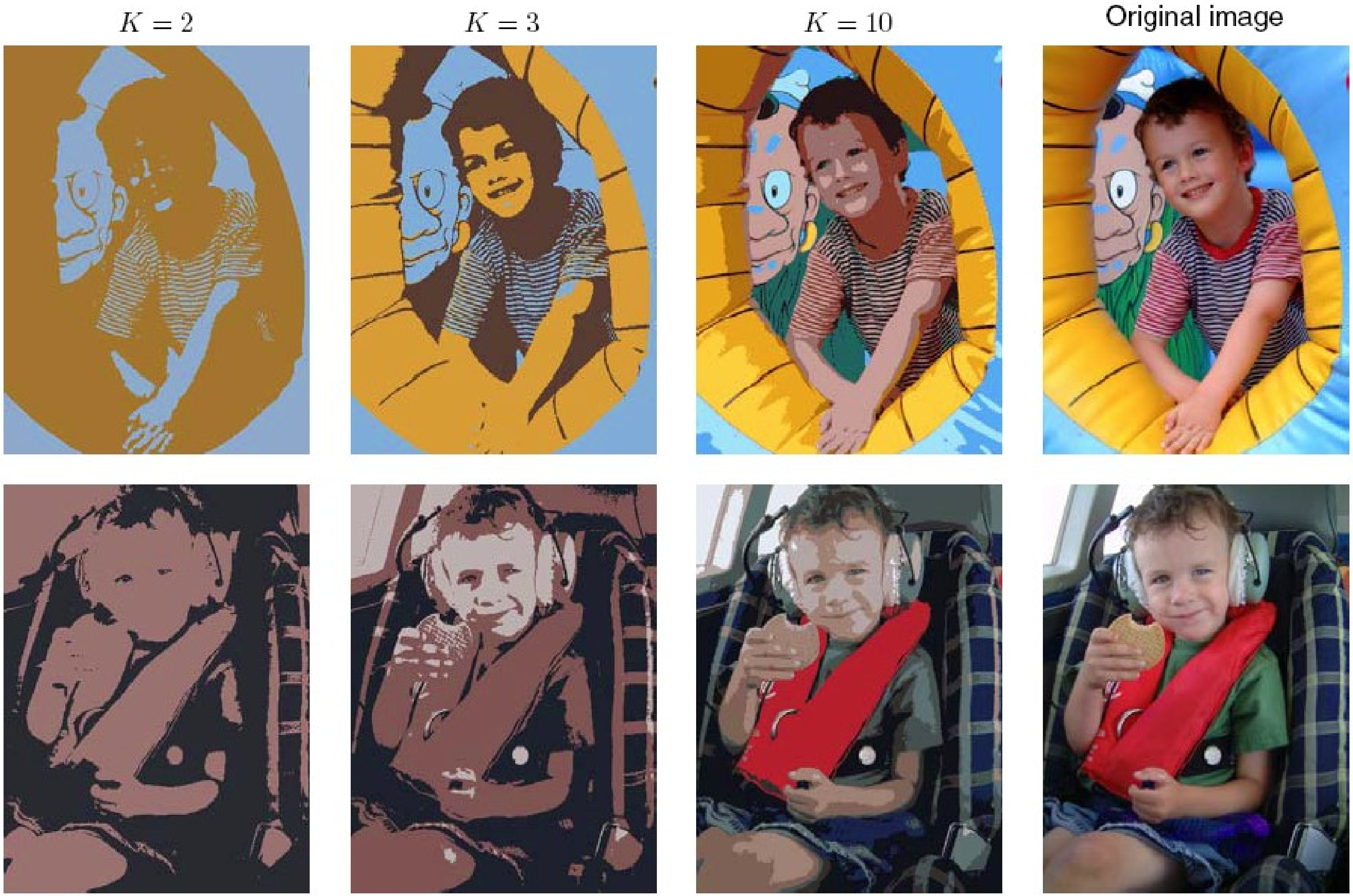


Update μ_k



K-means Clustering

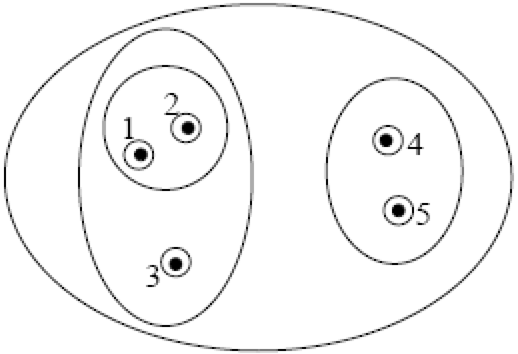
- Image Segmentation and Compression



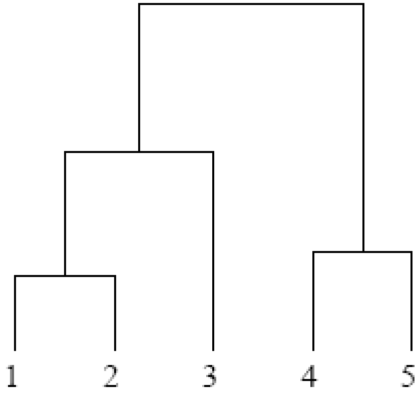
Hierarchical Clustering

- Finds clusters at all levels simultaneously

Data



Dendrogram



- Agglomerative clustering: Bottom-up clustering

- Divisive clustering: Top-down clustering

Hierarchical Clustering

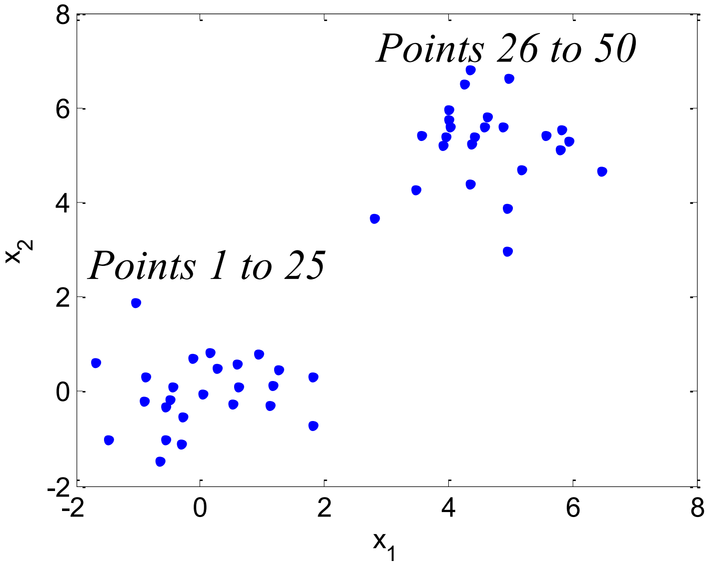
- Agglomerative Clustering Algorithm
  1. Start with each point in one cluster
  2. Merge the 2 closest clusters together
  3. Go back to step 2 until all points are in a single cluster

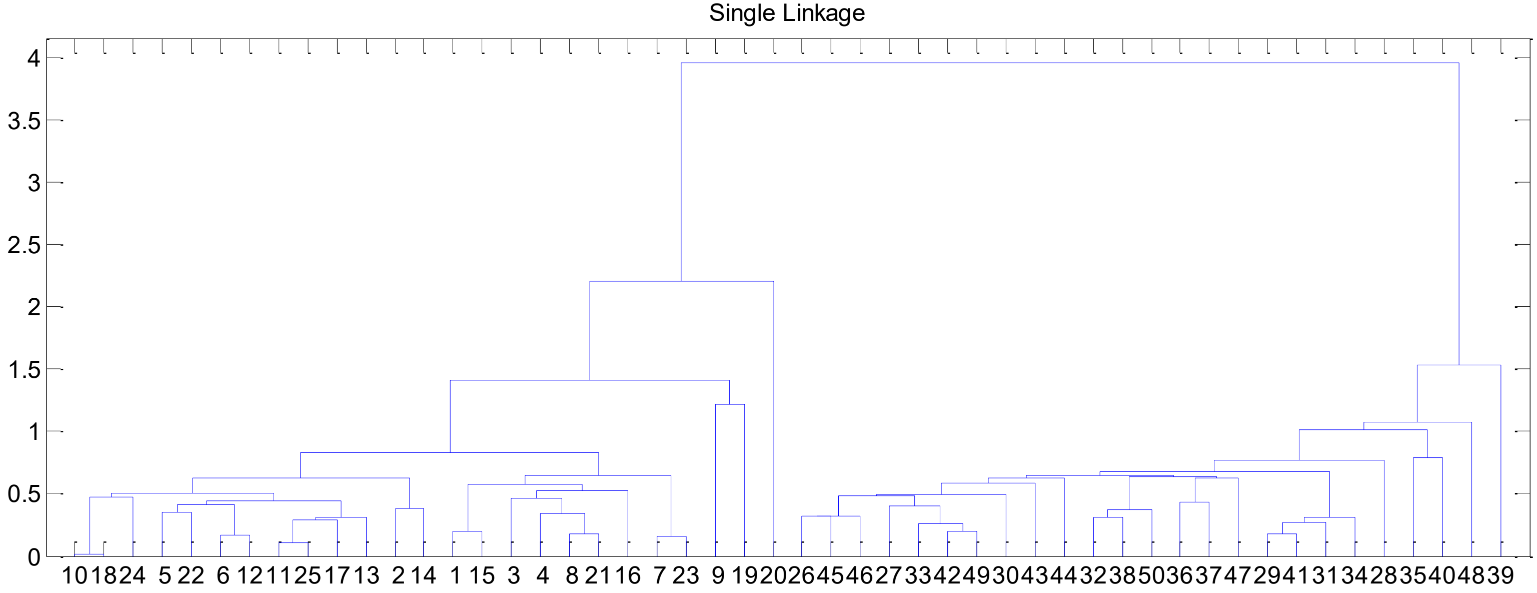
- How to define distance D between 2 clusters?
  - Single Linkage: $D(C_i, C_j) = \min_{\mathbf{x}_i \in C_i, \mathbf{x}_j \in C_j} d(\mathbf{x}_i, \mathbf{x}_j)$
  - Complete Linkage: $D(C_i, C_j) = \max_{\mathbf{x}_i \in C_i, \mathbf{x}_j \in C_j} d(\mathbf{x}_i, \mathbf{x}_j)$
  - Average Linkage: $D(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{\mathbf{x}_i \in C_i, \mathbf{x}_j \in C_j} d(\mathbf{x}_i, \mathbf{x}_j)$

where $d(\mathbf{x}_i, \mathbf{x}_j) = 1 - \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 \big/ 2\sigma^2\right)$

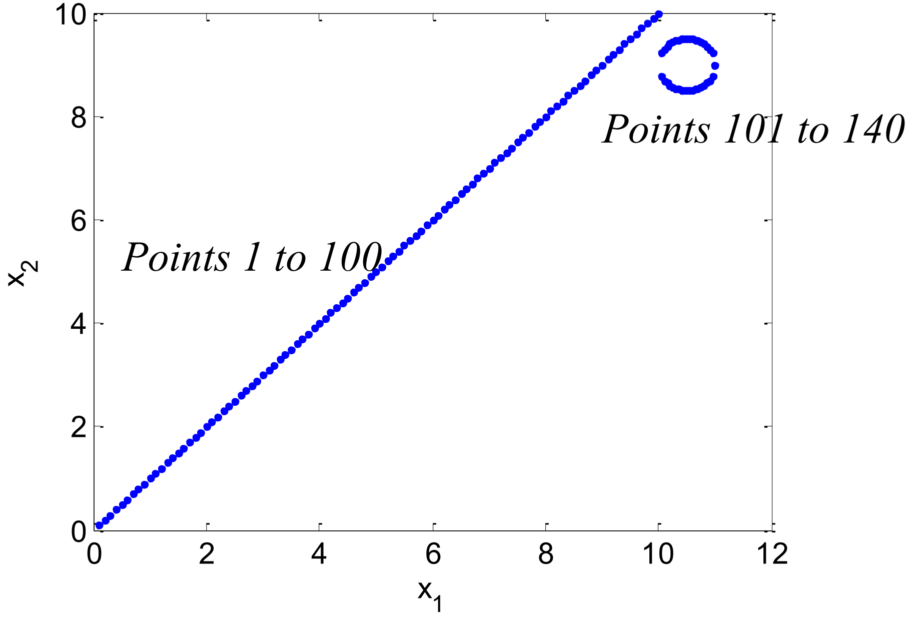
Or $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|$

Hierarchical Clustering

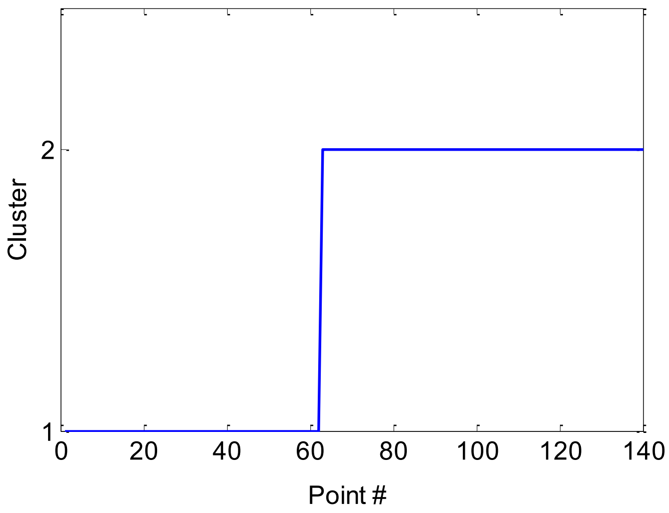




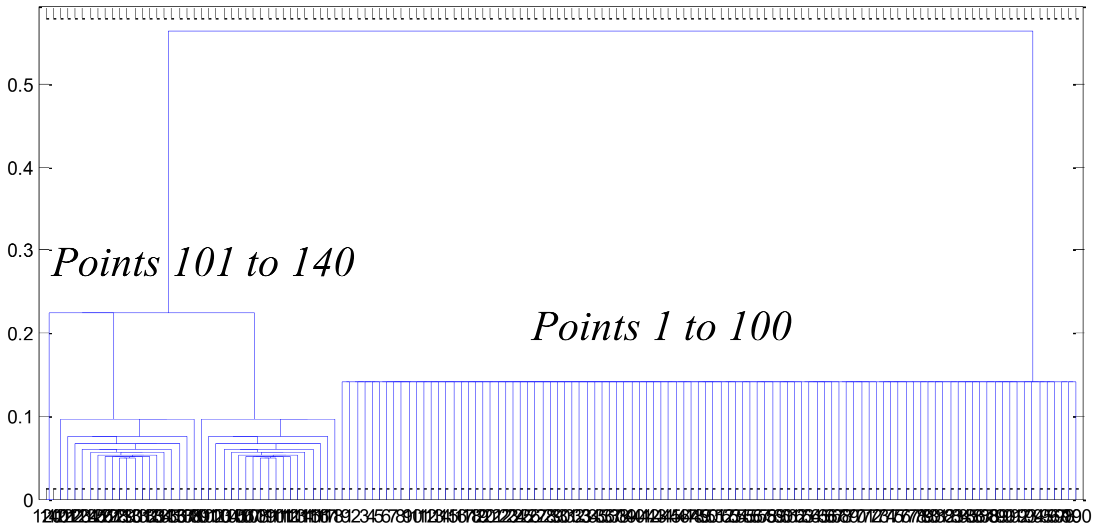
Hierarchical Clustering

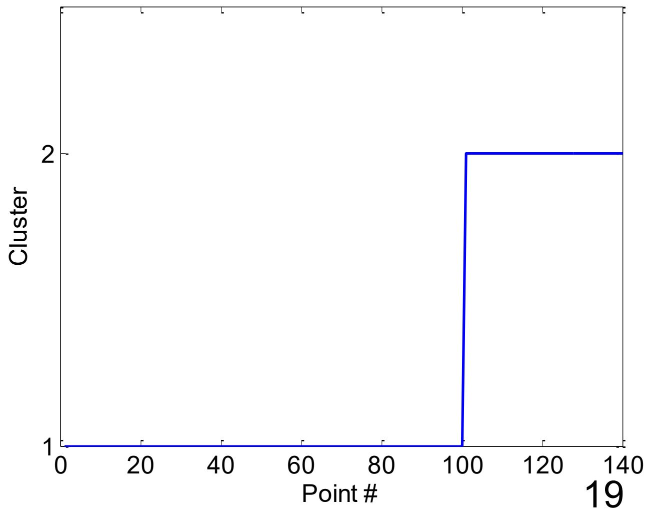


K-means



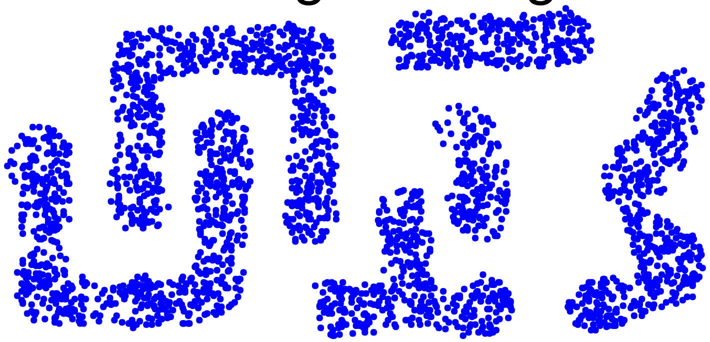
Single Linkage Dendrogram



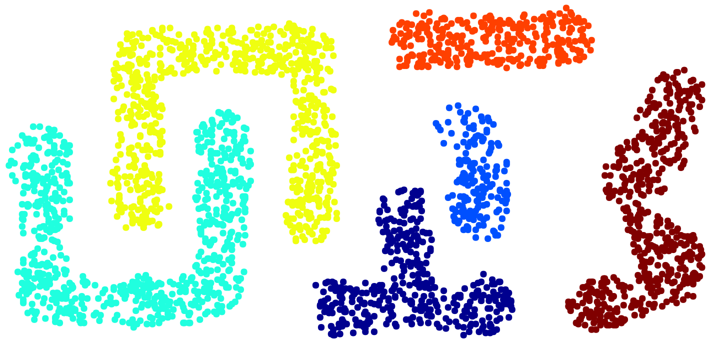


Hierarchical Clustering

- Advantage of Single Linkage: Can handle non-elliptical clusters

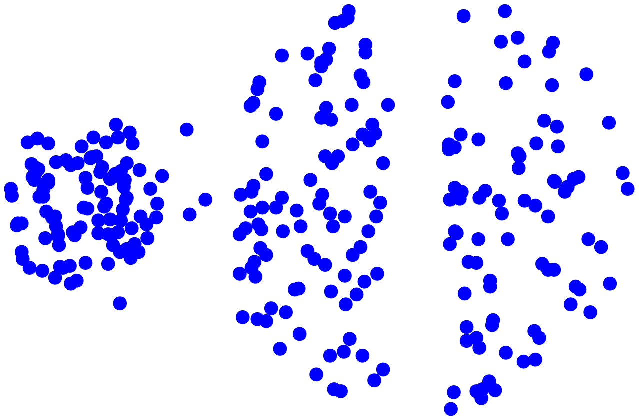


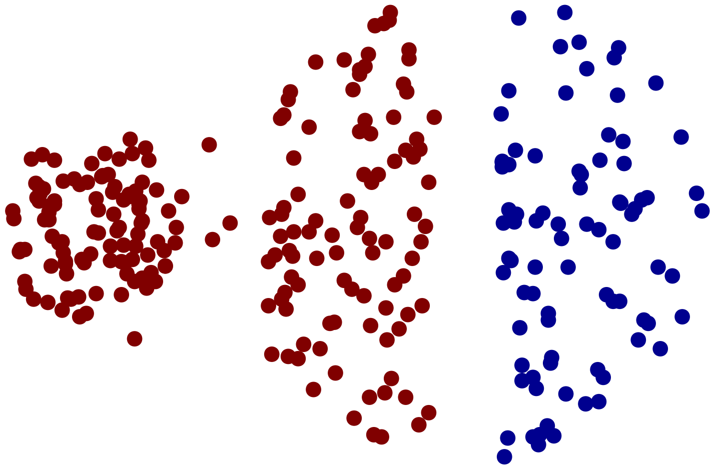
Original Points



Six Clusters

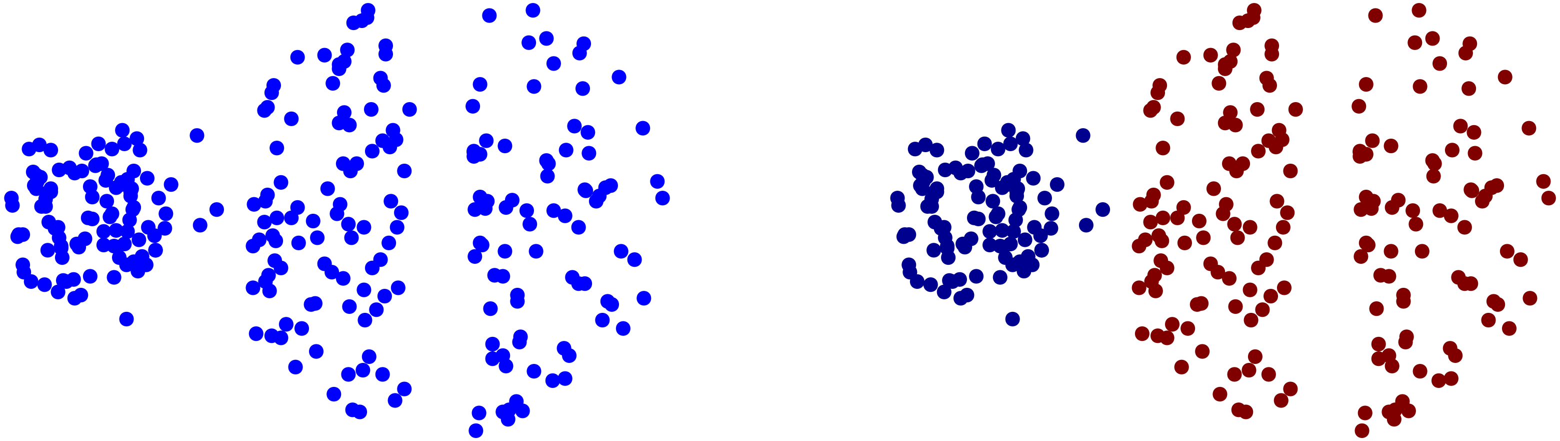
- Disadvantage: Sensitive to noise





Hierarchical Clustering

- Advantage of Complete Linkage: Less susceptible to noise



- Disadvantage: Tends to break large clusters and biased towards globular clusters