Web Mining

Cluster I (K-means, K-means++)

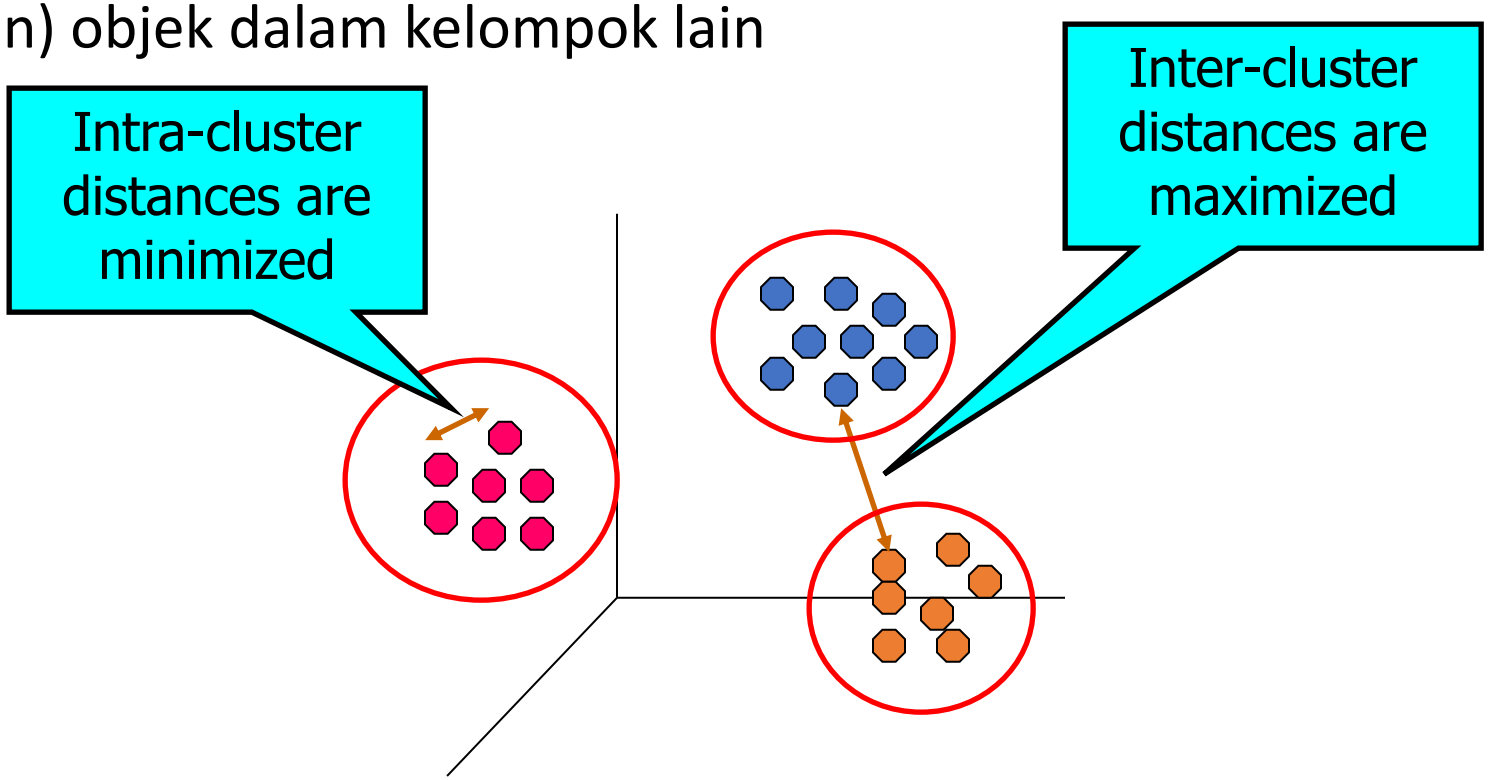
Prodi Teknik Informatika

Universitas Trunojoyo Madura

2024

What is Cluster Analysis?

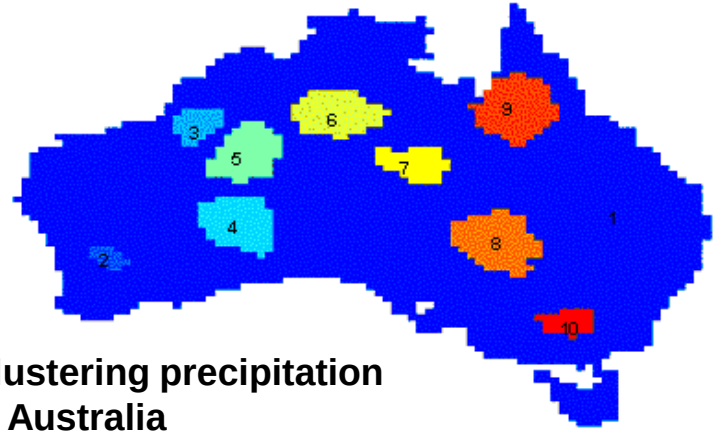
- Menemukan kelompok objek sedemikian rupa sehingga objek dalam suatu kelompok akan serupa (atau terkait) satu sama lain dan berbeda dari (atau tidak terkait dengan) objek dalam kelompok lain



Applications of Cluster Analysis

- **Understanding**
  - Kelompokkan dokumen terkait untuk penelusuran, kelompok gen dan protein yang memiliki fungsi serupa, atau kelompok saham dengan fluktuasi harga yang serupa
- **Summarization**
  - Kurangi ukuran data set yang besar

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP



Clustering precipitation in Australia

What is not Cluster Analysis?

- Simple segmentation
  - Membagi siswa ke dalam kelompok pendaftaran yang berbeda menurut abjad, berdasarkan nama belakang
- Results of a query
  - Groupings adalah hasil dari spesifikasi eksternal
  - Clustering adalah pengelompokan objek berdasarkan data
- Supervised classification
  - Memiliki informasi label kelas
- Association Analysis
  - Koneksi lokal vs. global

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

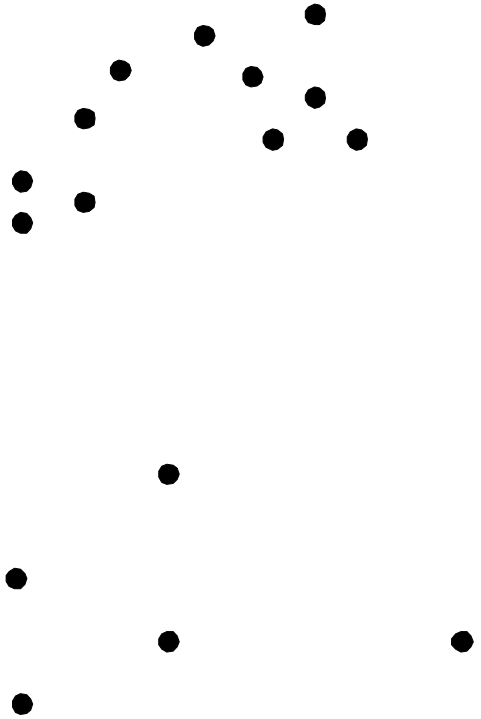
Two Clusters

Four Clusters

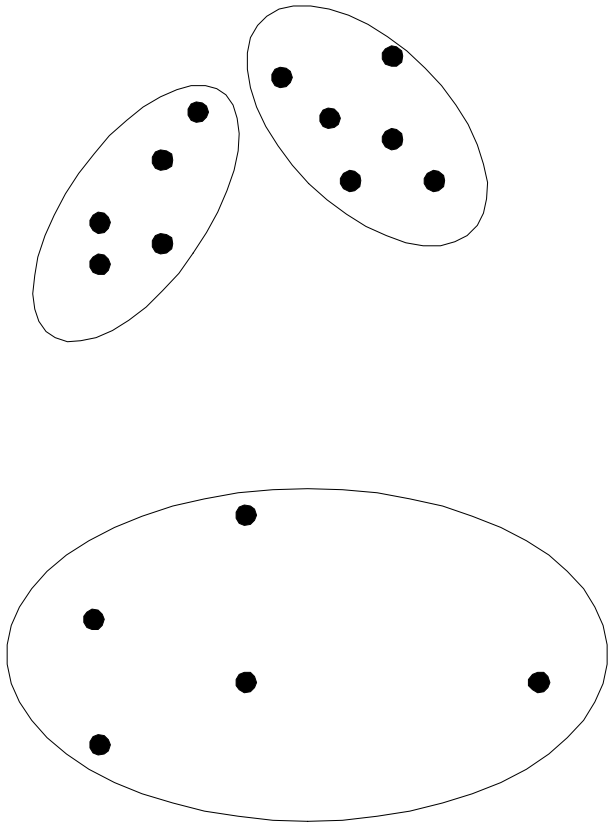
Types of Clusterings

- Pengklusteran adalah sekumpulan kluster
- Perbedaan penting antara kumpulan kluster hierarkis dan partisi
- Partitional Clustering
  - Pembagian objek data menjadi subset (cluster) yang tidak tumpang tindih sehingga setiap objek data berada dalam satu subset
- Hierarchical clustering
  - Sekumpulan kluster berjenjang yang diatur sebagai hierarki hierarkis

Partitional Clustering

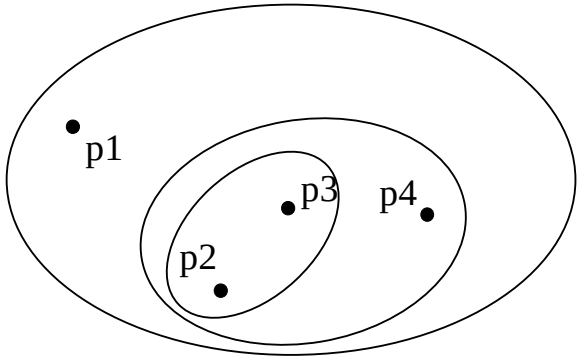


Original Points

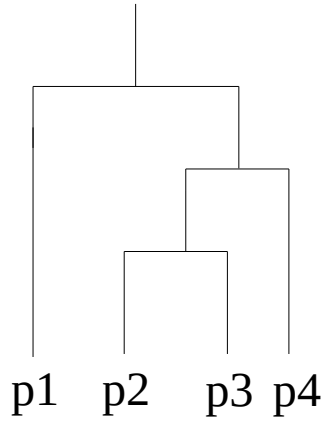


A Partitional Clustering

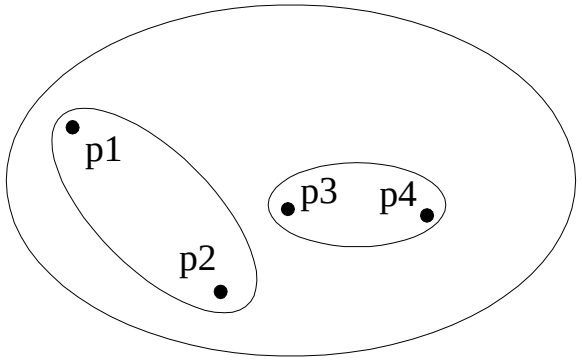
Hierarchical Clustering



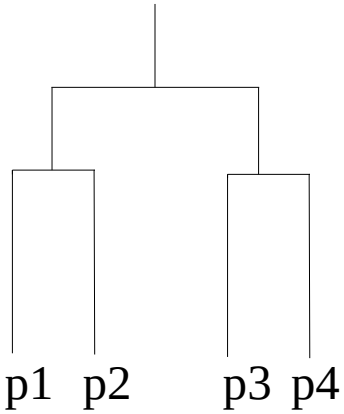
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Other Distinctions Between Sets of Clusters

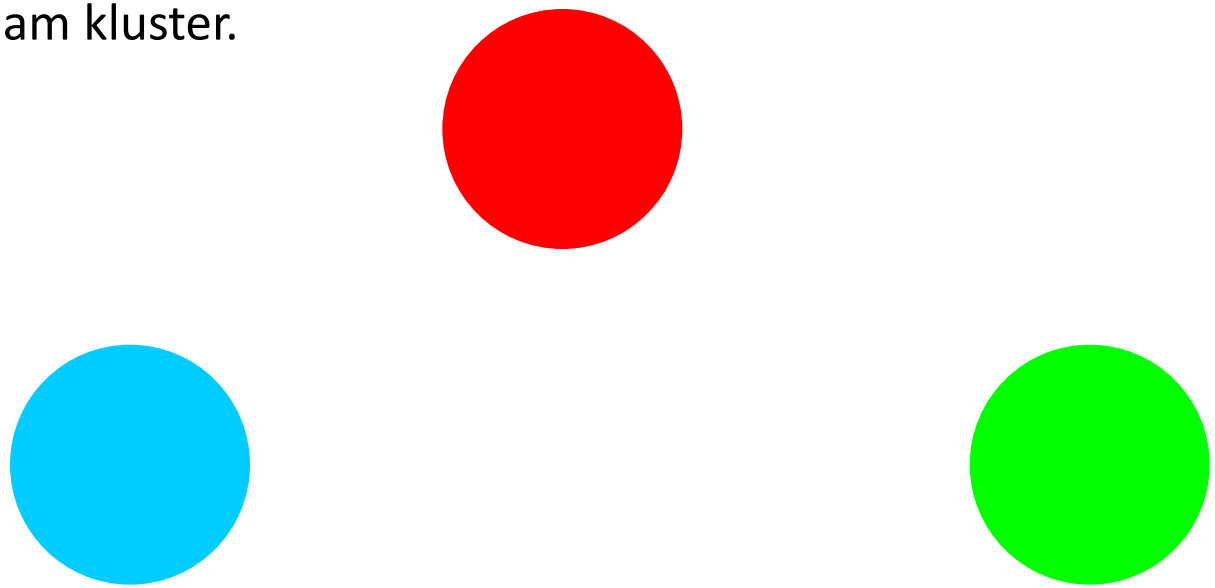
- Exclusive versus non-exclusive
  - Dalam pengelompokan non-eksklusif, banyak titik mungkin termasuk dalam beberapa kluster.
  - Dapat mewakili beberapa kelas atau titik 'batas’
- Fuzzy versus non-fuzzy
  - Dalam pengelompokan fuzzy, sebuah titik milik setiap kelompok dengan bobot antara 0 dan 1
  - Bobot harus berjumlah 1
  - Pengelompokan probabilistik memiliki karakteristik yang serupa
- Partial versus complete
  - Dalam beberapa kasus, kami hanya ingin mengelompokkan beberapa data
- Heterogeneous versus homogeneous
  - Cluster dengan ukuran, bentuk, dan kepadatan yang sangat berbeda

Types of Clusters

- Kelompok yang terpisah dengan baik
- Kluster berbasis pusat
- Kluster yang berdekatan
- Kluster berbasis kepadatan
- Properti atau Konseptual
- Dijelaskan oleh Fungsi Objektif

Types of Clusters: Well-Separated

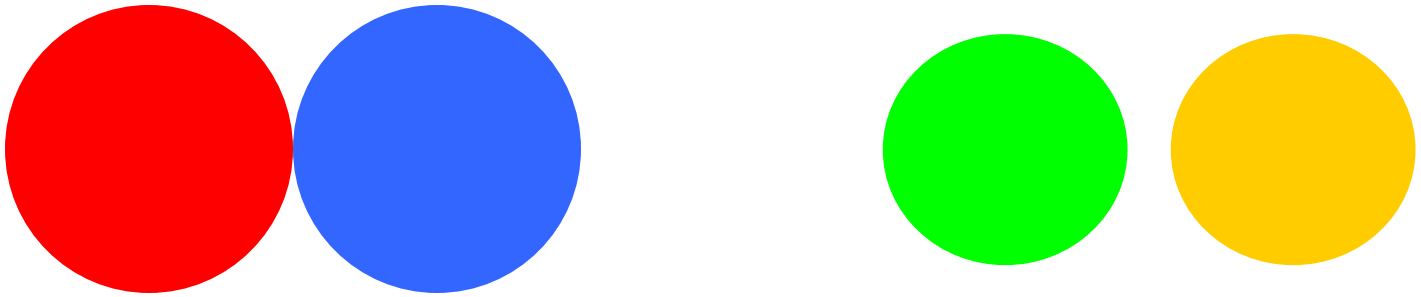
- Well-Separated Clusters:
  - Kluster adalah sekumpulan titik sedemikian rupa sehingga setiap titik dalam kluster lebih dekat (atau lebih mirip) ke setiap titik lain dalam kluster daripada ke titik mana pun yang tidak ada dalam kluster.



3 well-separated clusters

Types of Clusters: Center-Based

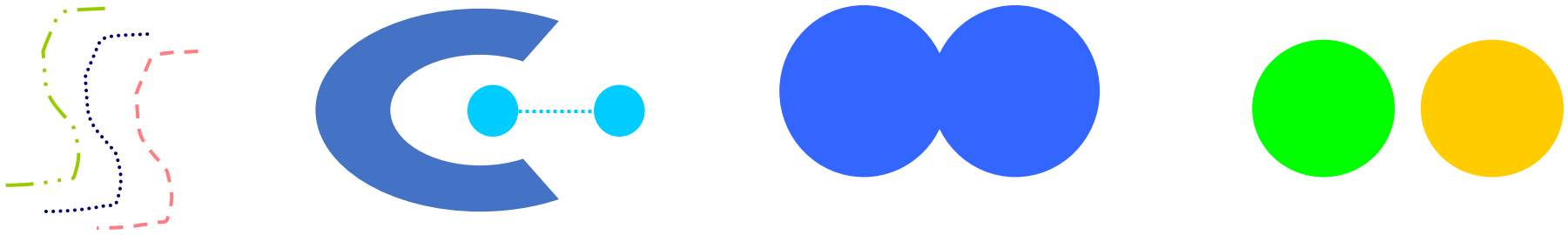
- Center-based
  - Cluster adalah sekumpulan objek sedemikian rupa sehingga objek dalam cluster lebih dekat (lebih mirip) ke "pusat" cluster, daripada ke pusat cluster lainnya
  - Pusat gugus seringkali berupa centroid, rata-rata dari semua titik dalam gugus, atau medoid, titik paling "representatif" dari sebuah kluster



4 center-based clusters

Types of Clusters: Contiguity-Based

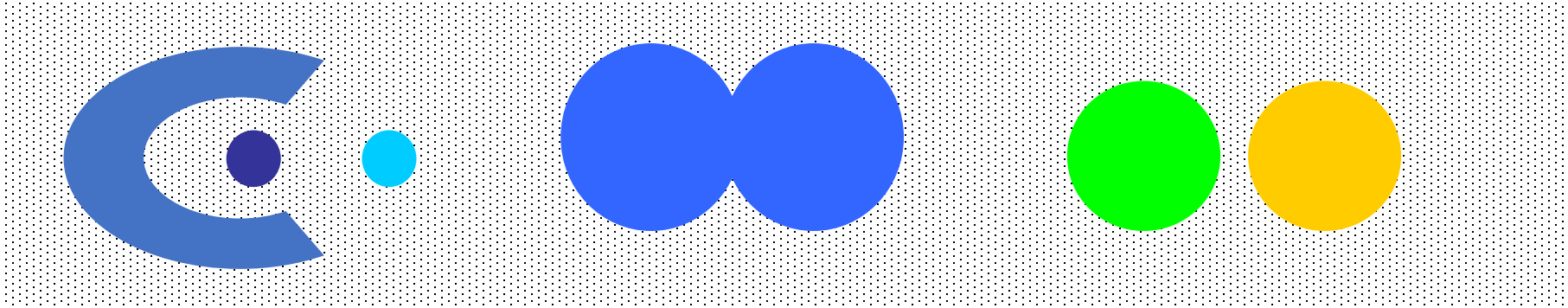
- Contiguous Cluster (Nearest neighbor or Transitive)
  - Kluster adalah sekumpulan titik sedemikian rupa sehingga titik dalam kluster lebih dekat (atau lebih mirip) ke satu atau lebih titik lain dalam kluster daripada titik mana pun yang tidak ada dalam kluster.



8 contiguous clusters

Types of Clusters: Density-Based

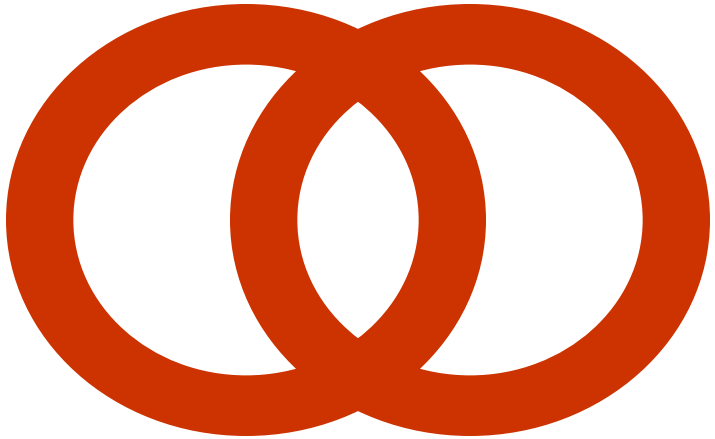
- Density-based
  - Cluster adalah wilayah titik padat, yang dipisahkan oleh wilayah kepadatan rendah, dari wilayah lain dengan kepadatan tinggi.
  - Digunakan ketika cluster tidak teratur atau terjalin, dan ketika ada kebisingan dan outlier.



6 density-based clusters

Types of Clusters: Conceptual Clusters

- Shared Property or Conceptual Clusters
  - Menemukan kluster yang berbagi beberapa properti umum atau mewakili konsep tertentu.



2 Overlapping Circles

Types of Clusters: Objective Function

- Clusters Defined by an Objective Function
  - Menemukan kluster yang meminimalkan atau memaksimalkan fungsi objektif.
  - Hitung semua cara yang mungkin untuk membagi titik-titik menjadi kelompok dan mengevaluasi 'kebaikan' dari setiap kumpulan potensial kelompok dengan menggunakan fungsi objektif yang diberikan. (NP Keras)
  - Dapat memiliki tujuan global atau lokal.
    - Algoritme pengelompokan hierarkis biasanya memiliki tujuan lokal
    - Algoritme partisi biasanya memiliki tujuan global
  - Variasi dari pendekatan fungsi objektif global adalah menyesuaikan data dengan model parameter.
    - Parameter untuk model ditentukan dari data.
    - Model campuran mengasumsikan bahwa data tersebut adalah 'campuran' dari sejumlah distribusi statistik.

Map Clustering Problem to a Different Problem

- Petakan masalah pengelompokan ke domain yang berbeda dan selesaikan masalah terkait di domain tersebut
  - Matriks kedekatan mendefinisikan grafik berbobot, di mana simpul adalah titik yang dikelompokkan, dan tepi tertimbang mewakili kedekatan antar titik
  - Pengklusteran setara dengan memecah grafik menjadi komponen yang terhubung, satu untuk setiap kluster.
  - Ingin meminimalkan bobot tepi antar kluster dan memaksimalkan bobot tepi dalam kluster

Characteristics of the Input Data Are Important

- Jenis pengukuran kedekatan atau kepadatan
  - Pusat pengelompokan
  - Tergantung pada data dan aplikasi
- Karakteristik data yang mempengaruhi kedekatan dan/atau kepadatan adalah
  - Dimensi
  - Jarang
  - Jenis atribut
  - Hubungan khusus dalam data
  - Misalnya, autokorelasi
  - Distribusi data
- Noise and Outliers
  - Sering mengganggu pengoperasian algoritma pengelompokan

Clustering Algorithms

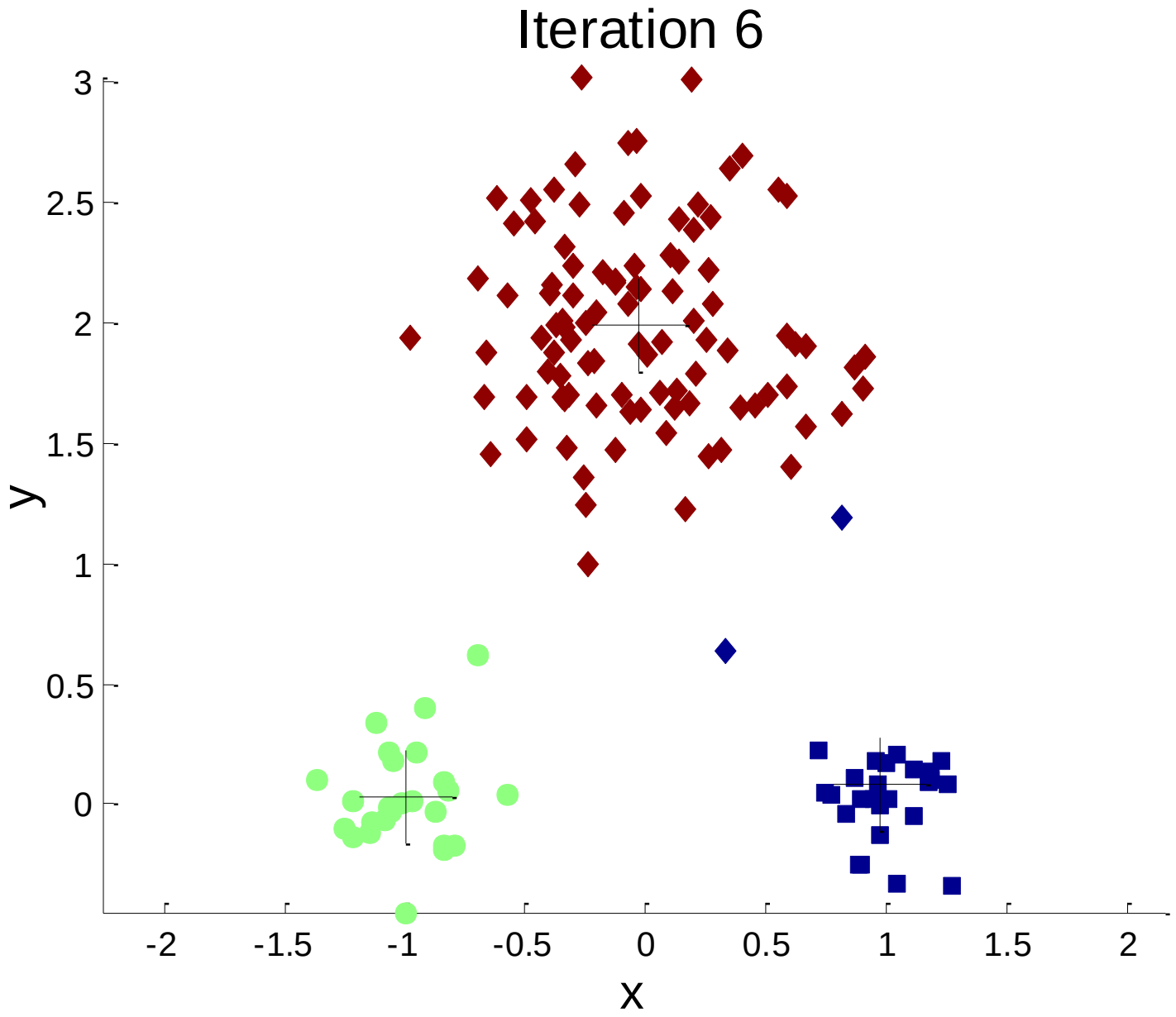
- K-means and its variants
- Hierarchical clustering
- Density-based clustering

K-means Clustering

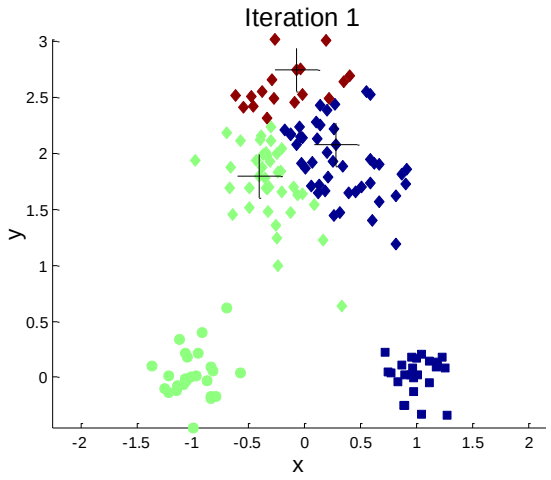
- Pendekatan pengelompokan partisi
- Jumlah kluster, K, harus ditentukan
- Setiap cluster dikaitkan dengan centroid (titik tengah)
- Setiap titik ditetapkan ke gugus dengan centroid terdekat
- Algoritma dasarnya sangat sederhana

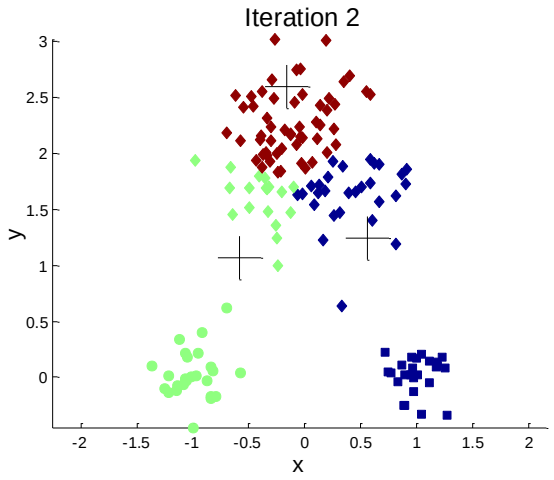
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

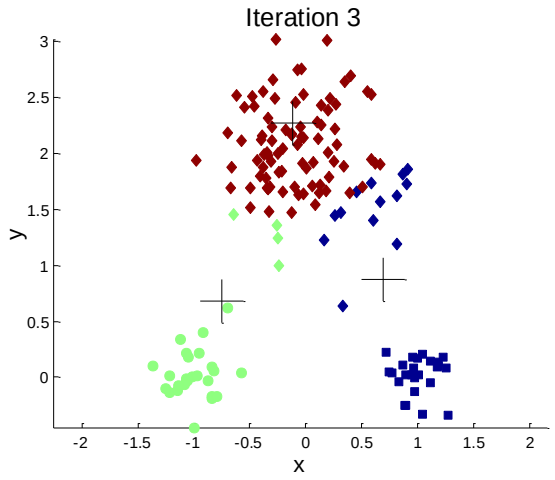
Example of K-means Clustering

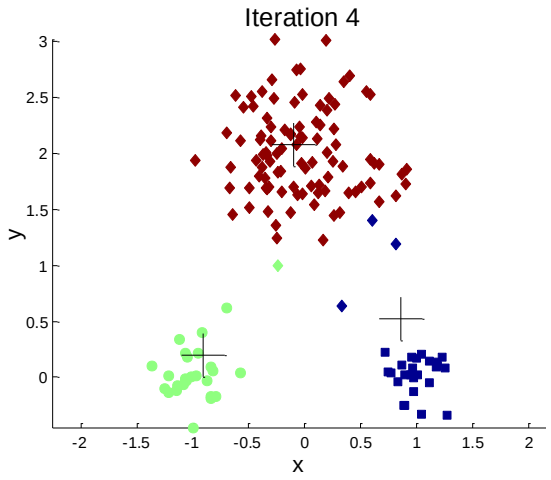


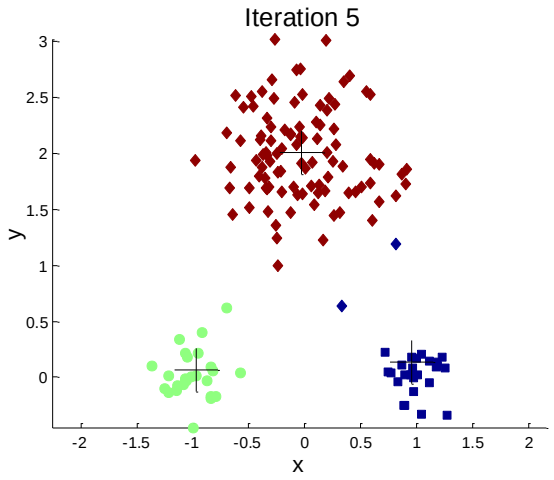
Example of K-means Clustering

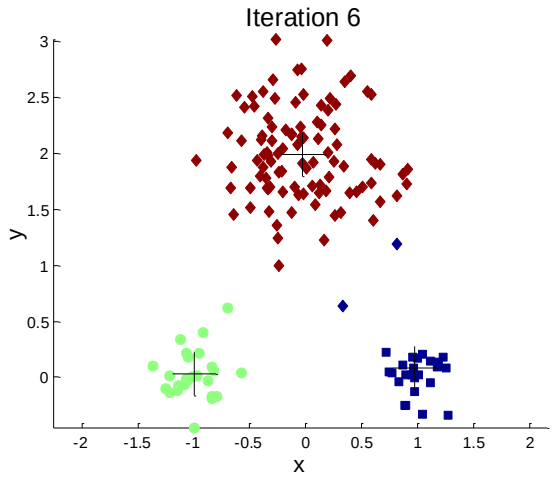












K-means Clustering – Details

- Centroid awal sering dipilih secara acak.
  - Cluster yang dihasilkan bervariasi dari satu run ke run lainnya.
- Sentroid (biasanya) adalah rata-rata titik-titik dalam kluster.
- 'Kedekatan' diukur dengan jarak Euclidean, kesamaan kosinus, korelasi, dll.
- K-means akan menyatu untuk ukuran kesamaan umum yang disebutkan di atas.
- Sebagian besar konvergensi terjadi dalam beberapa iterasi pertama.
  - Seringkali kondisi berhenti diubah menjadi 'Sampai relatif sedikit titik berubah cluster'
- Complexity is O(n * K * I * d)
  - n = number of points, K = number of clusters,
    I = number of iterations, d = number of attributes

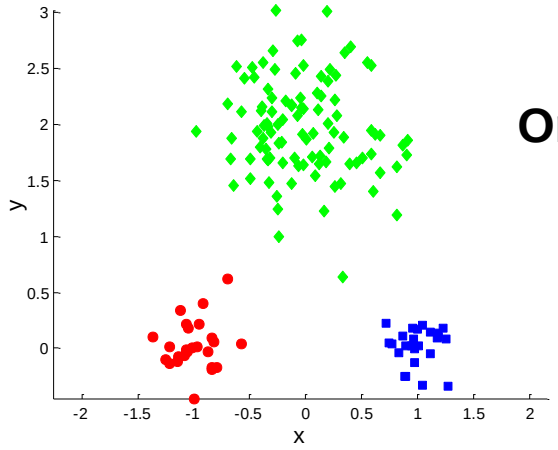
Evaluating K-means Clusters

- Ukuran yang paling umum adalah Sum of Squared Error (SSE)
  - Untuk setiap titik, kesalahannya adalah jarak ke gugus terdekat
  - Untuk mendapatkan SSE, kami kuadratkan kesalahan ini dan menjumlahkannya.

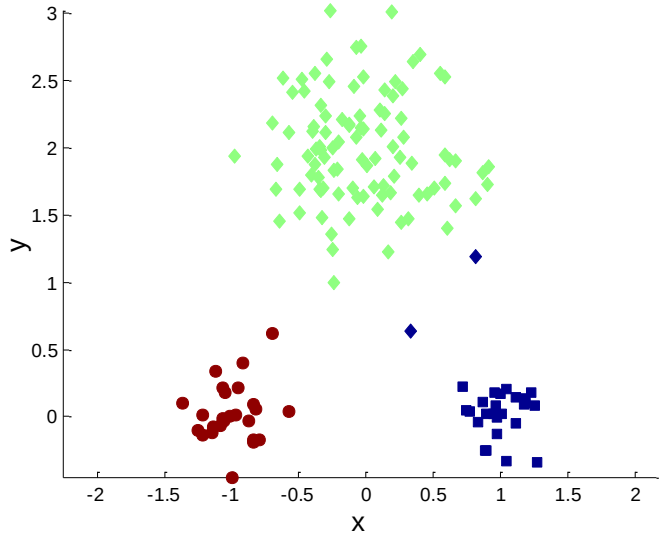
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x adalah titik data dalam klaster Ci dan mi adalah titik representatif untuk klaster Ci
    - dapat menunjukkan bahwa mi sesuai dengan pusat (rata-rata) gugus
  - Mengingat dua set kluster, kami lebih memilih yang memiliki kesalahan terkecil
  - Salah satu cara mudah untuk mengurangi SSE adalah dengan menambah K, jumlah klaster
    - Pengelompokan yang baik dengan K yang lebih kecil dapat memiliki SSE yang lebih rendah daripada pengelompokan yang buruk dengan K yang lebih tinggi

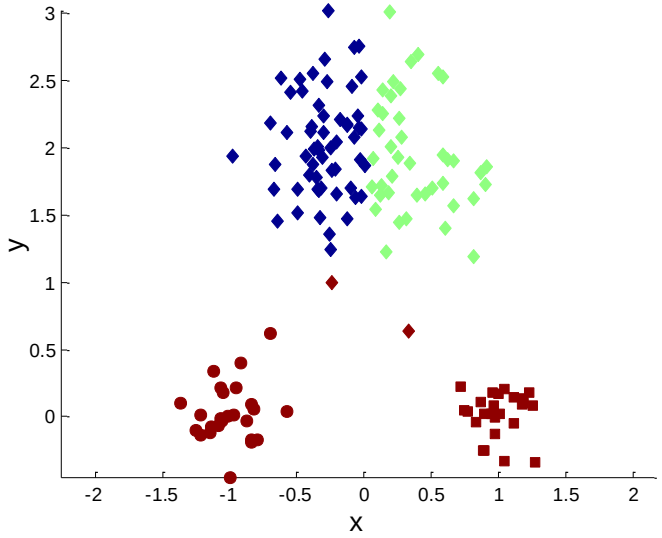
Two different K-means Clusterings



Original Points



Optimal Clustering



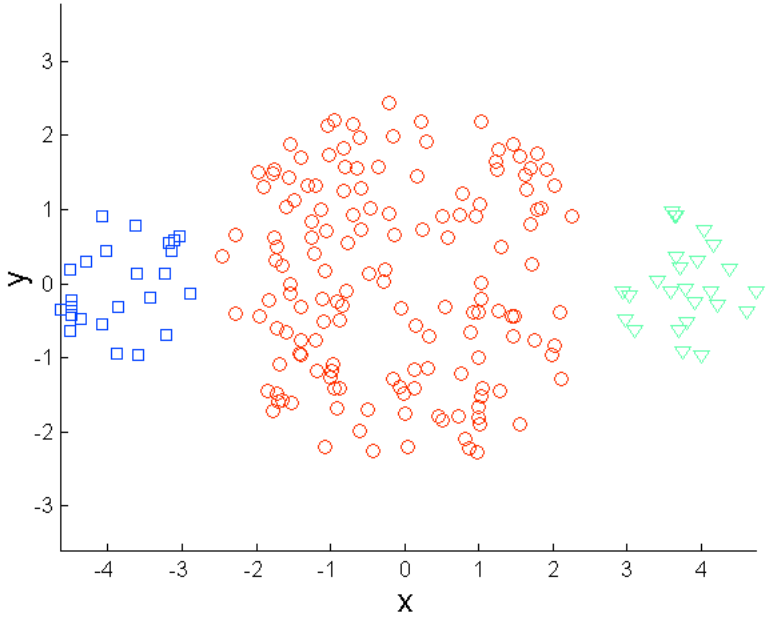
Sub-optimal Clustering

Limitations of K-means

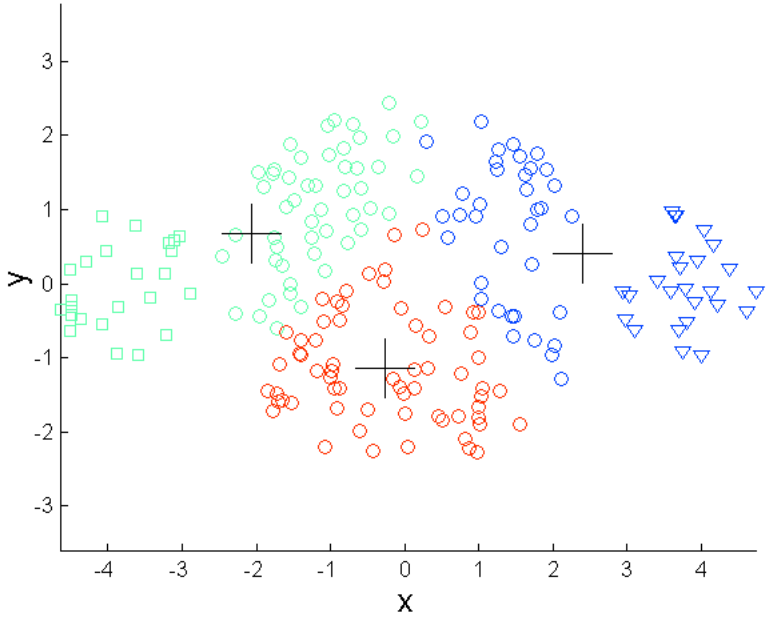
- K-means memiliki masalah ketika cluster berbeda
  - Ukuran
  - Kepadatan
  - Bentuk non-bola

- K-means memiliki masalah ketika data berisi outlier.

Limitations of K-means: Differing Sizes

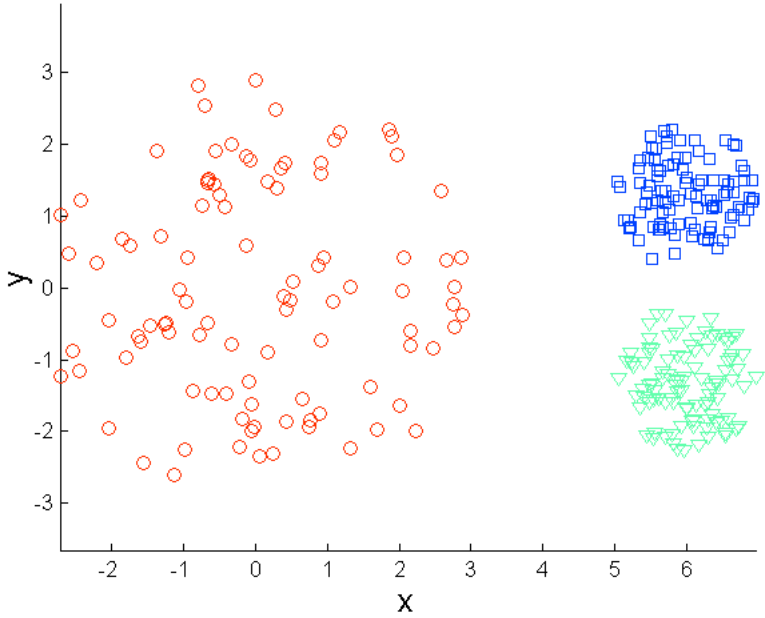


Original Points

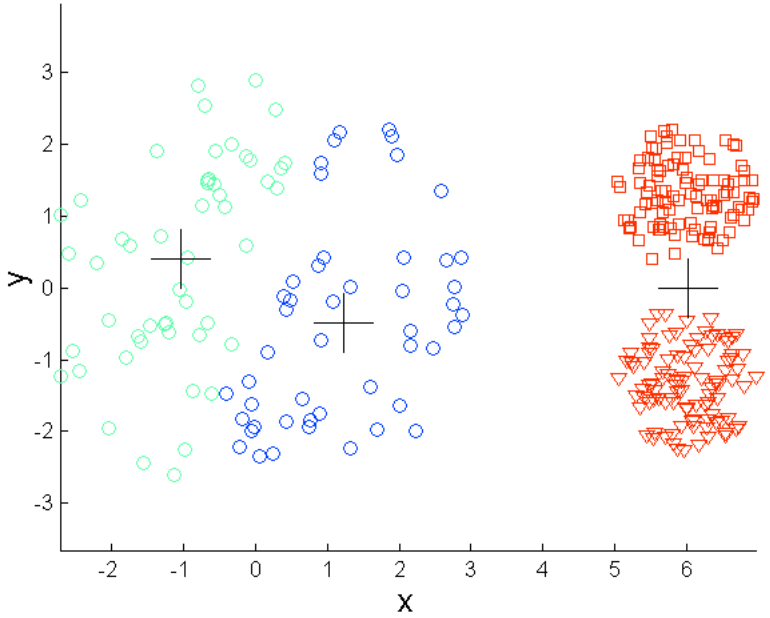


K-means (3 Clusters)

Limitations of K-means: Differing Density

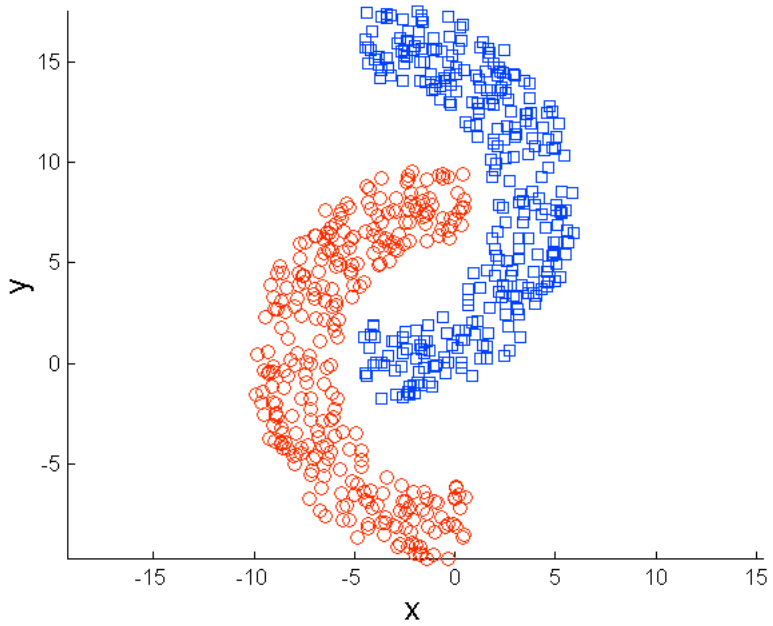


Original Points

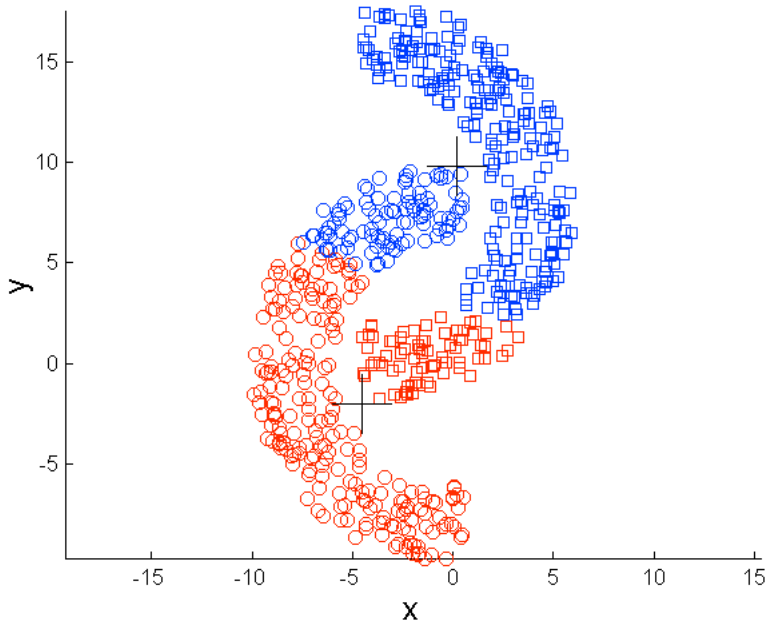


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

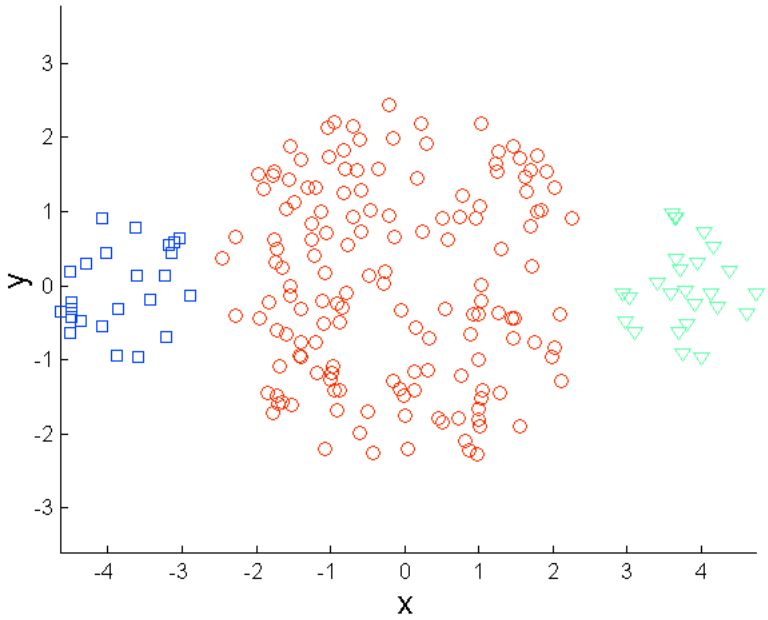


Original Points

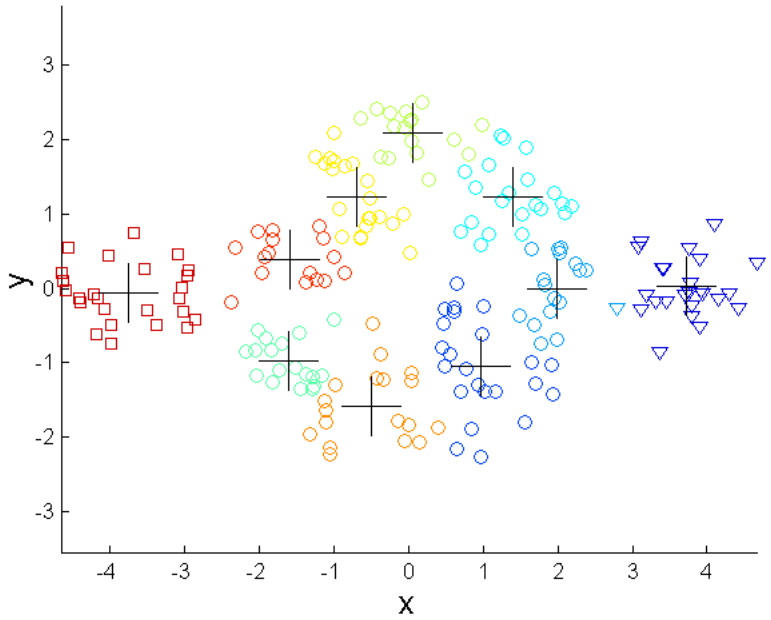


K-means (2 Clusters)

Overcoming K-means Limitations



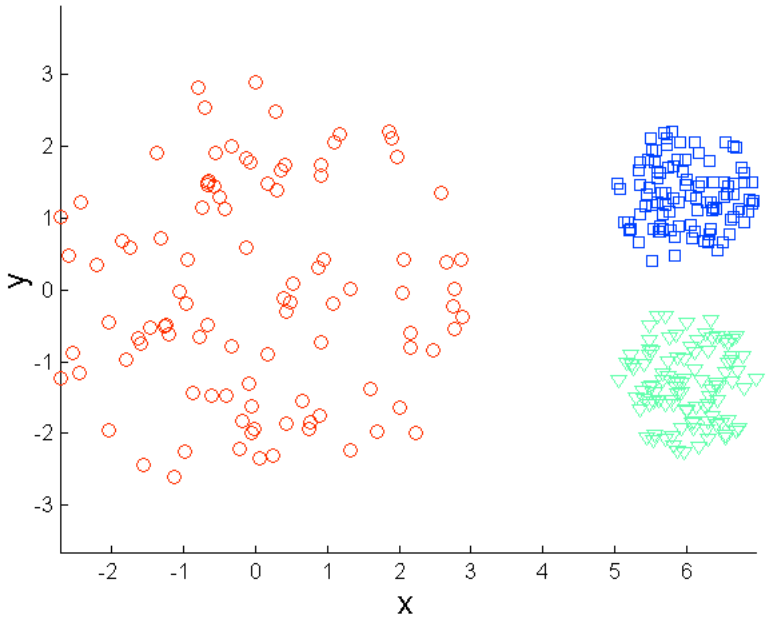
Original Points



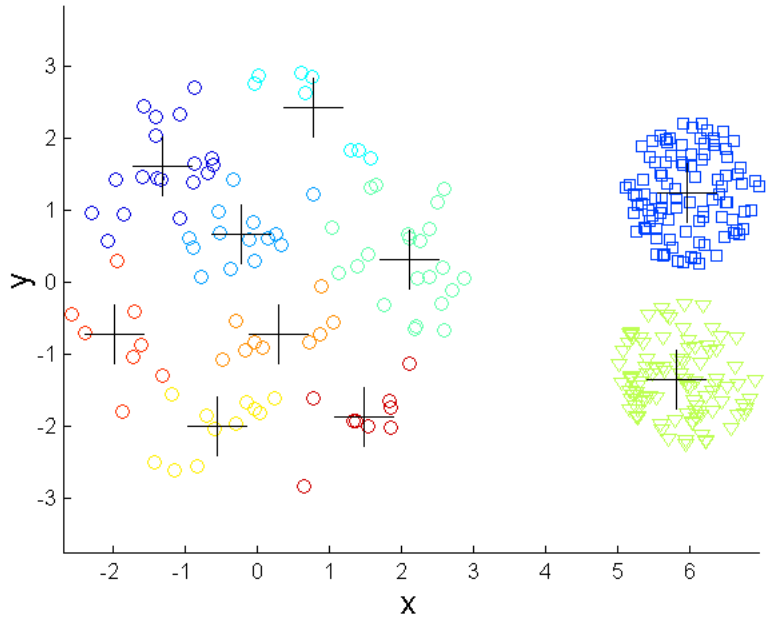
K-means Clusters

- Salah satu solusinya adalah dengan menggunakan banyak cluster.
- Temukan bagian dari cluster, tetapi perlu disatukan.

Overcoming K-means Limitations

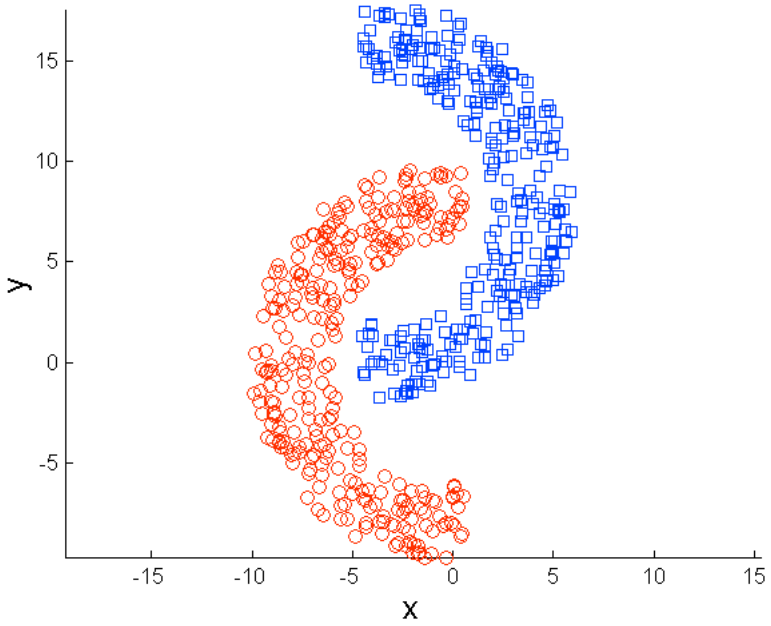


Original Points

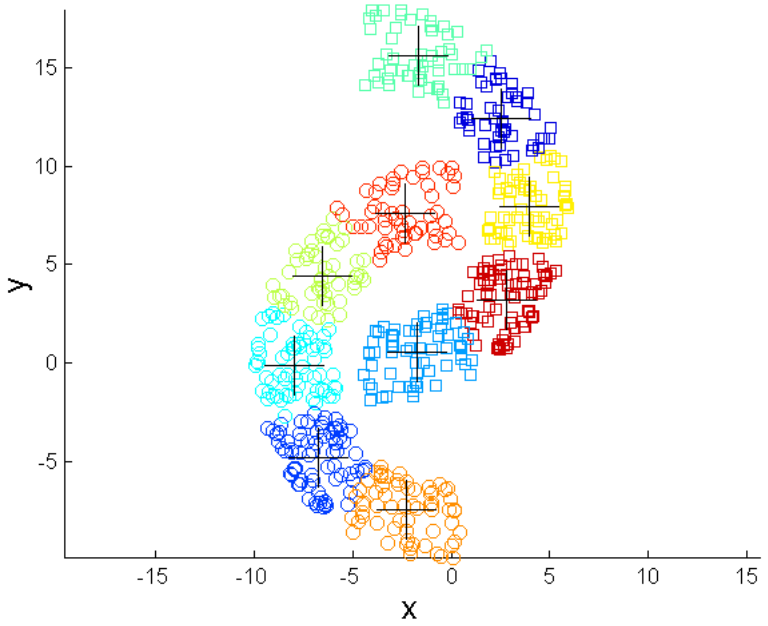


K-means Clusters

Overcoming K-means Limitations

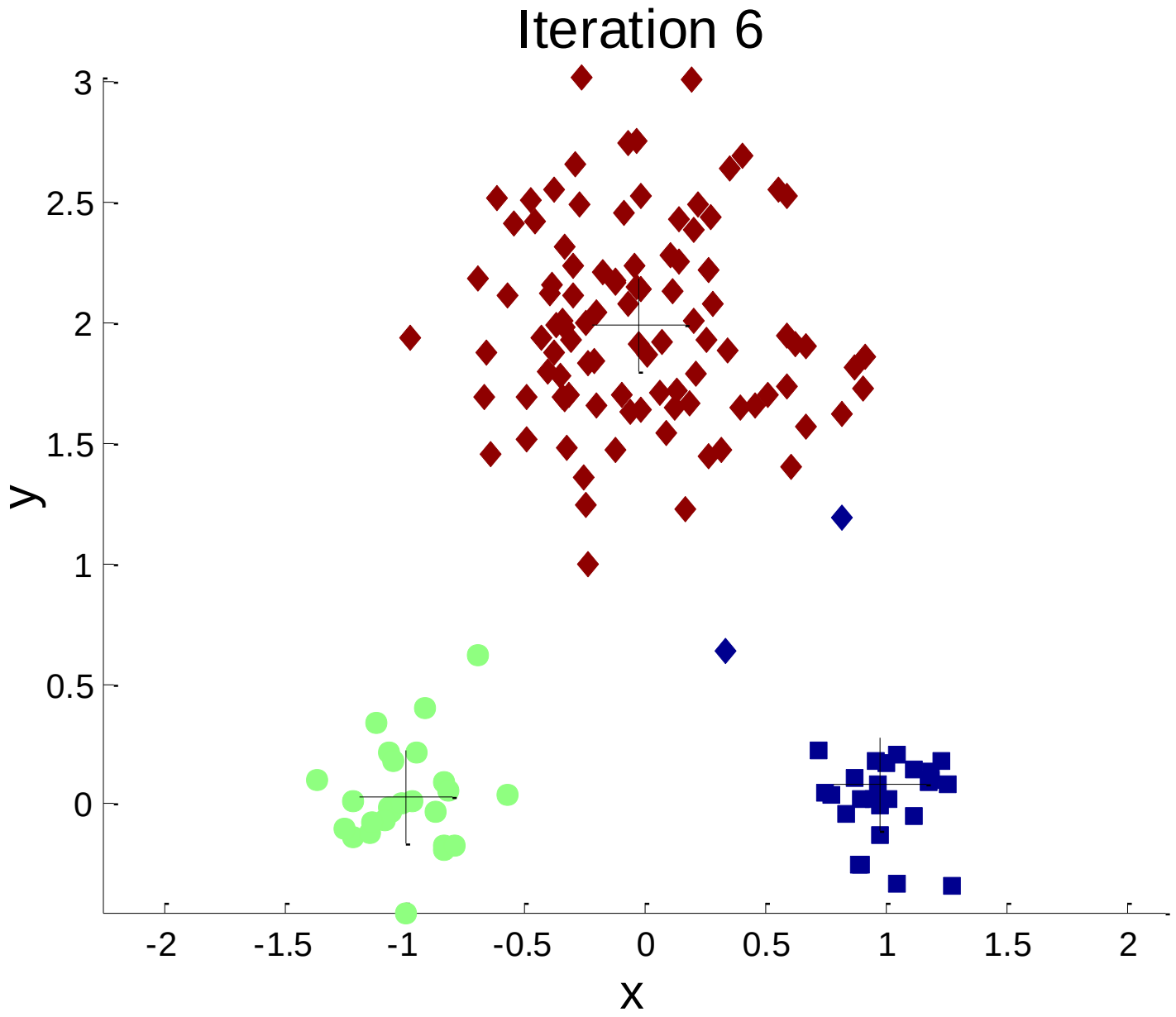


Original Points

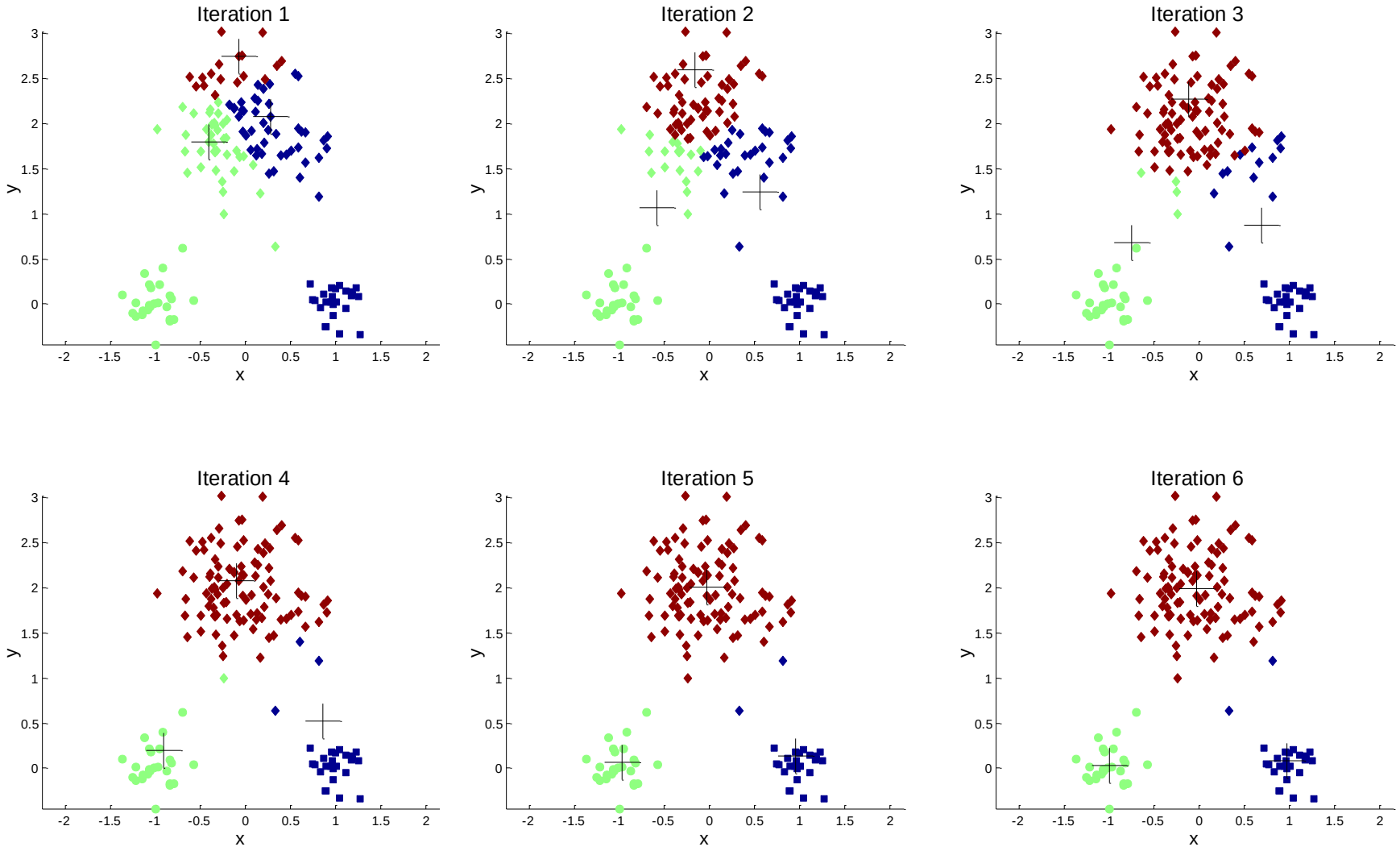


K-means Clusters

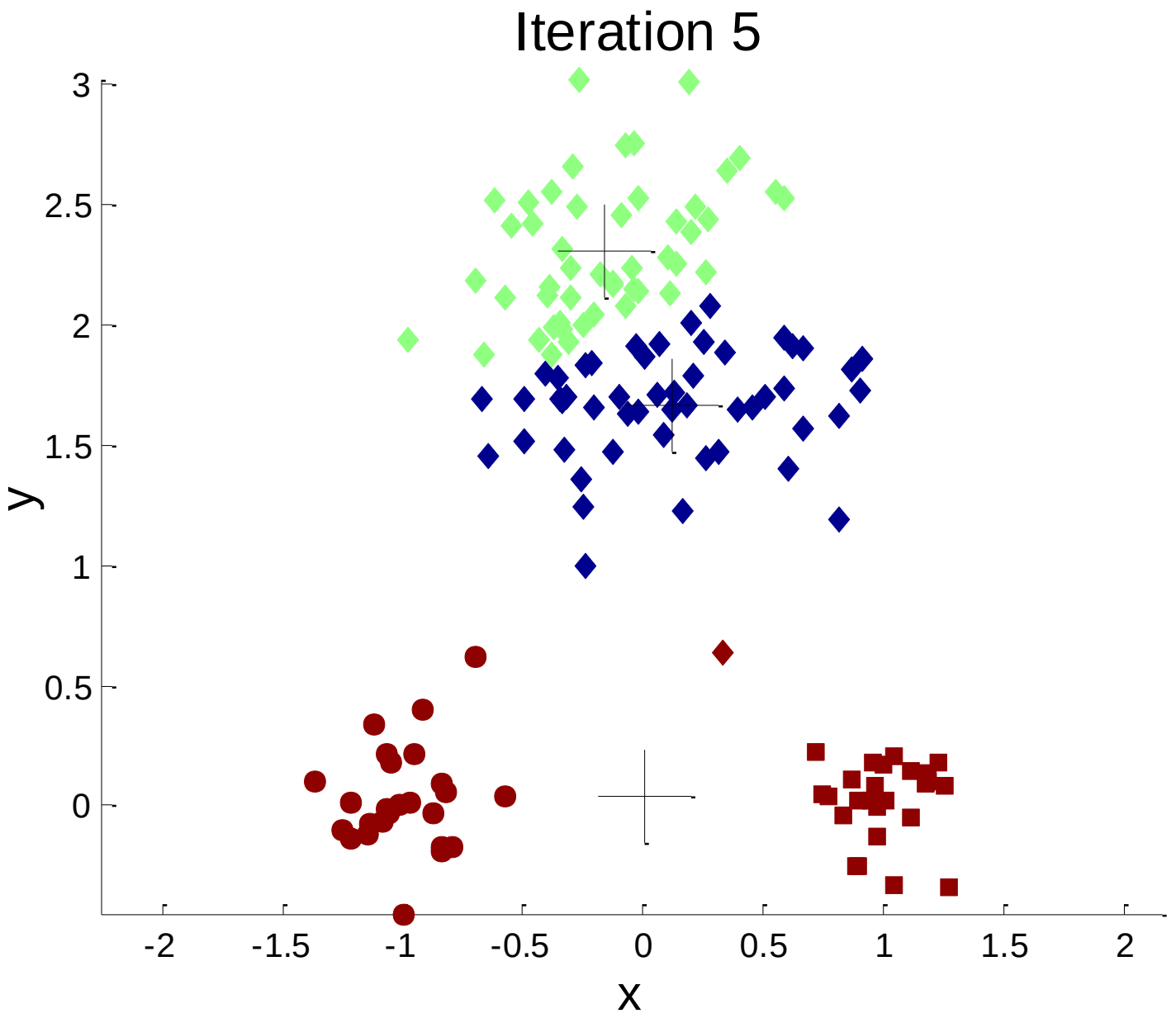
Importance of Choosing Initial Centroids



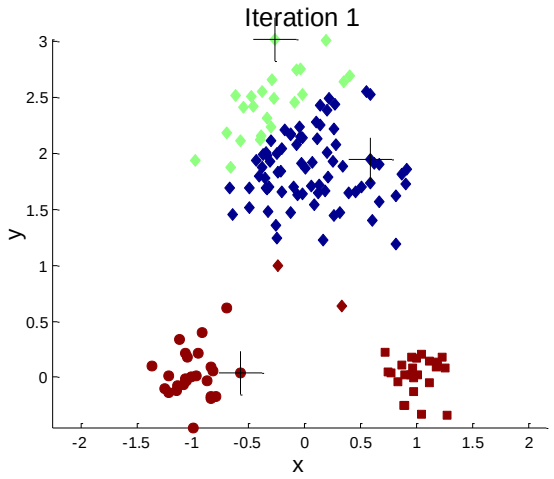
Importance of Choosing Initial Centroids

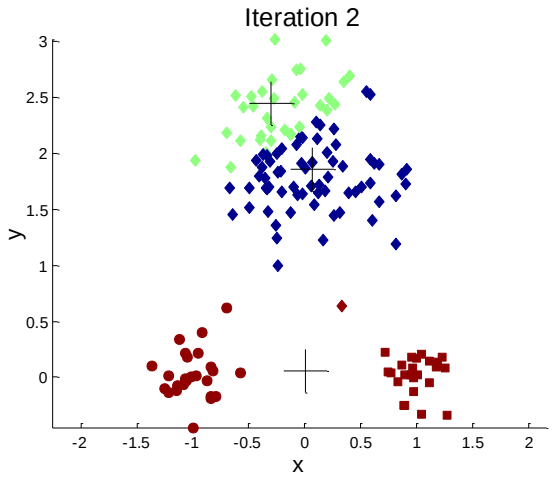


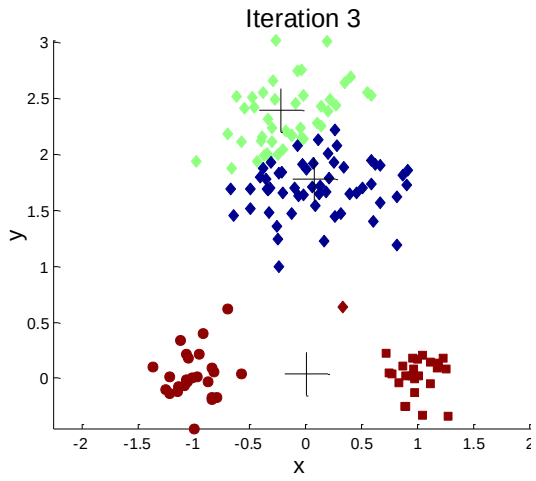
Importance of Choosing Initial Centroids …

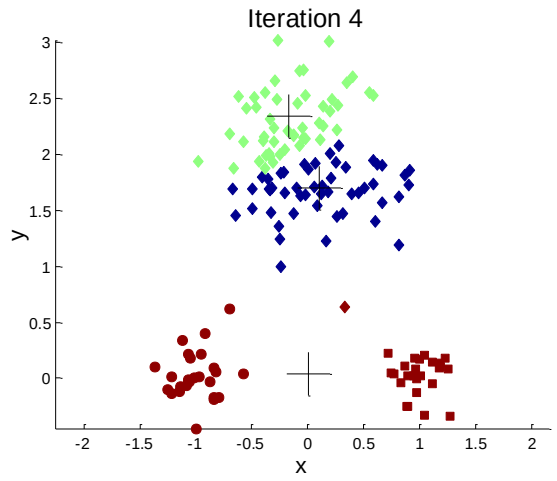


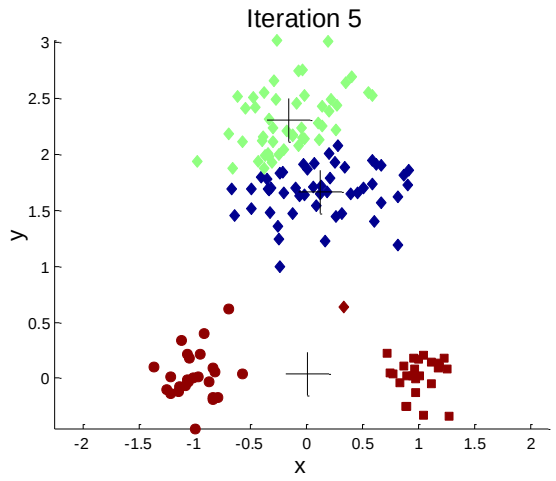
Importance of Choosing Initial Centroids …







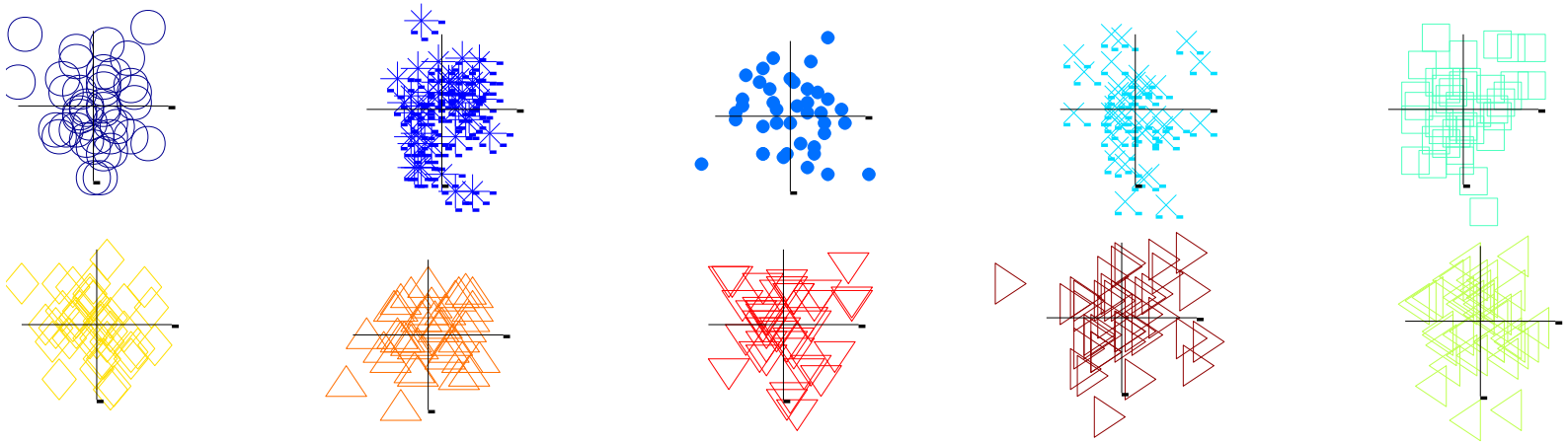




Problems with Selecting Initial Points

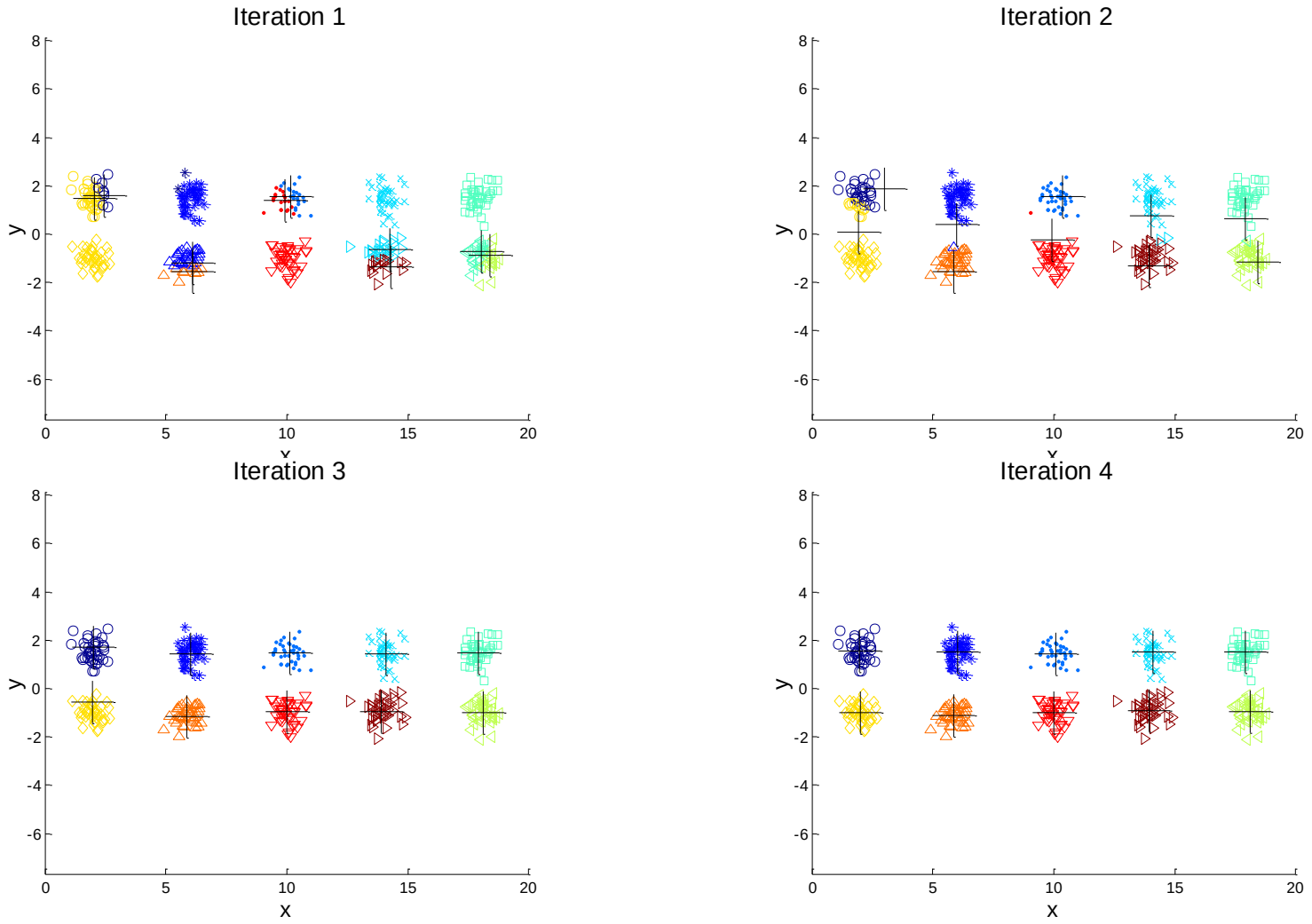
- Jika ada K kluster 'nyata' maka peluang untuk memilih satu centroid dari setiap cluster kecil.
  - Peluang relatif kecil ketika K besar
  - Jika cluster berukuran sama, n, maka
  -
$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K! n^K}{(Kn)^K} = \frac{K!}{K^K}$$
  - Misalnya, jika K = 10, maka probabilitas = 10!/1010 = 0,00036
  - Terkadang centroid awal akan menyesuaikan diri dengan cara yang 'benar', dan terkadang tidak
  - Pertimbangkan contoh lima pasang kluster

10 Clusters Example



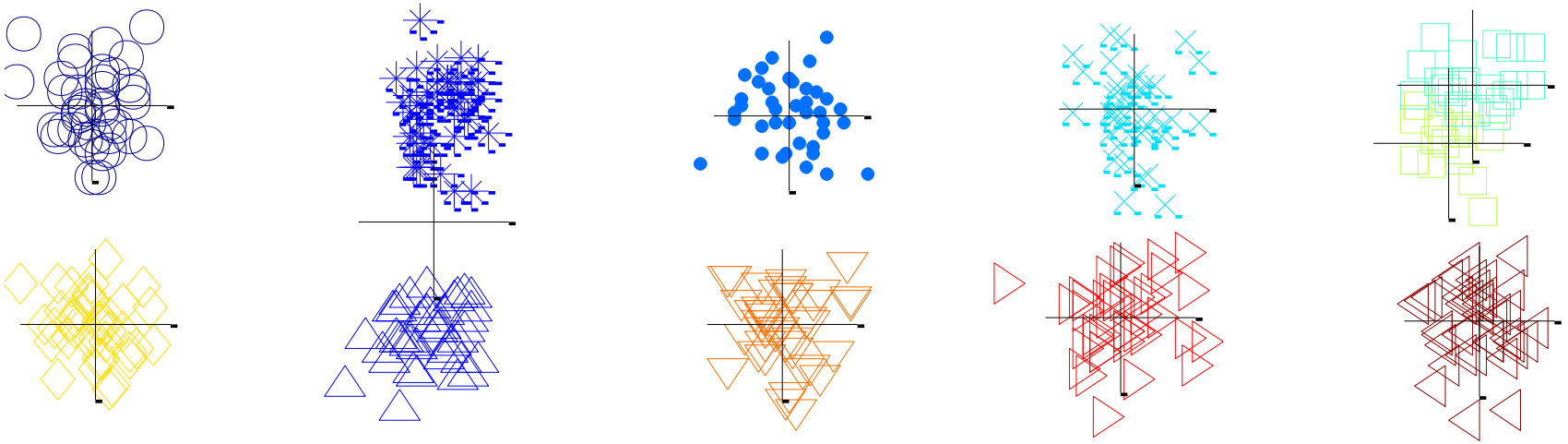
Starting with two initial centroids in one cluster of each pair of clusters

10 Clusters Example



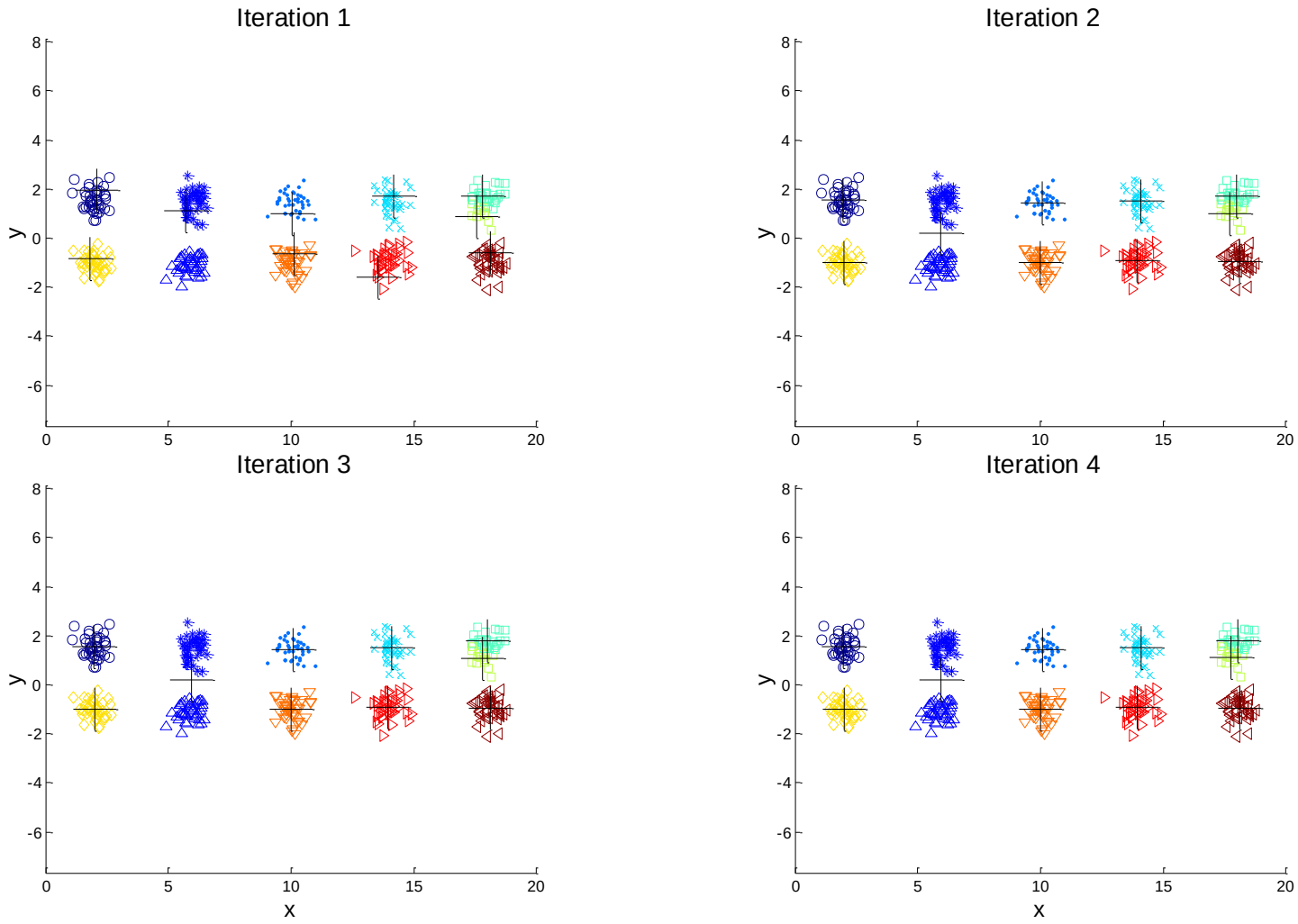
Starting with two initial centroids in one cluster of each pair of clusters

10 Clusters Example



Starting with some pairs of clusters having three initial centroids, while other have only one.

10 Clusters Example



Starting with some pairs of clusters having three initial centroids, while other have only one.

Solutions to Initial Centroids Problem

- Multiple runs
  - Membantu, tetapi probabilitas tidak ada di pihak Anda
- Sampel dan gunakan pengelompokan hierarkis untuk menentukan centroid awal
- Pilih lebih dari k centroid awal dan kemudian pilih di antara centroid awal ini
  - Pilih yang paling terpisah
- Postprocessing
- Hasilkan lebih banyak kluster lalu lakukan pengelompokan hierarkis
- Bisecting K-means
  - Tidak rentan terhadap masalah inisialisasi

K-means++

- Pendekatan ini bisa lebih lambat daripada inisialisasi acak, tetapi sangat konsisten menghasilkan hasil yang lebih baik dalam hal SSE
  - Algoritma k-means++ menjamin rasio perkiraan O(log k) dalam ekspektasi, di mana k adalah jumlah pusat
- Untuk memilih satu set centroid awal, C, lakukan hal berikut
- Select an initial point at random to be the first centroid

For k – 1 steps

Untuk masing-masing titik N, xi, 1 ≤ i ≤ N, temukan kuadrat minimum jarak ke centroid yang dipilih saat ini, C1, …, Cj, 1 ≤ j < k,
i.e., $\min_j$ d2(Cj, xi)

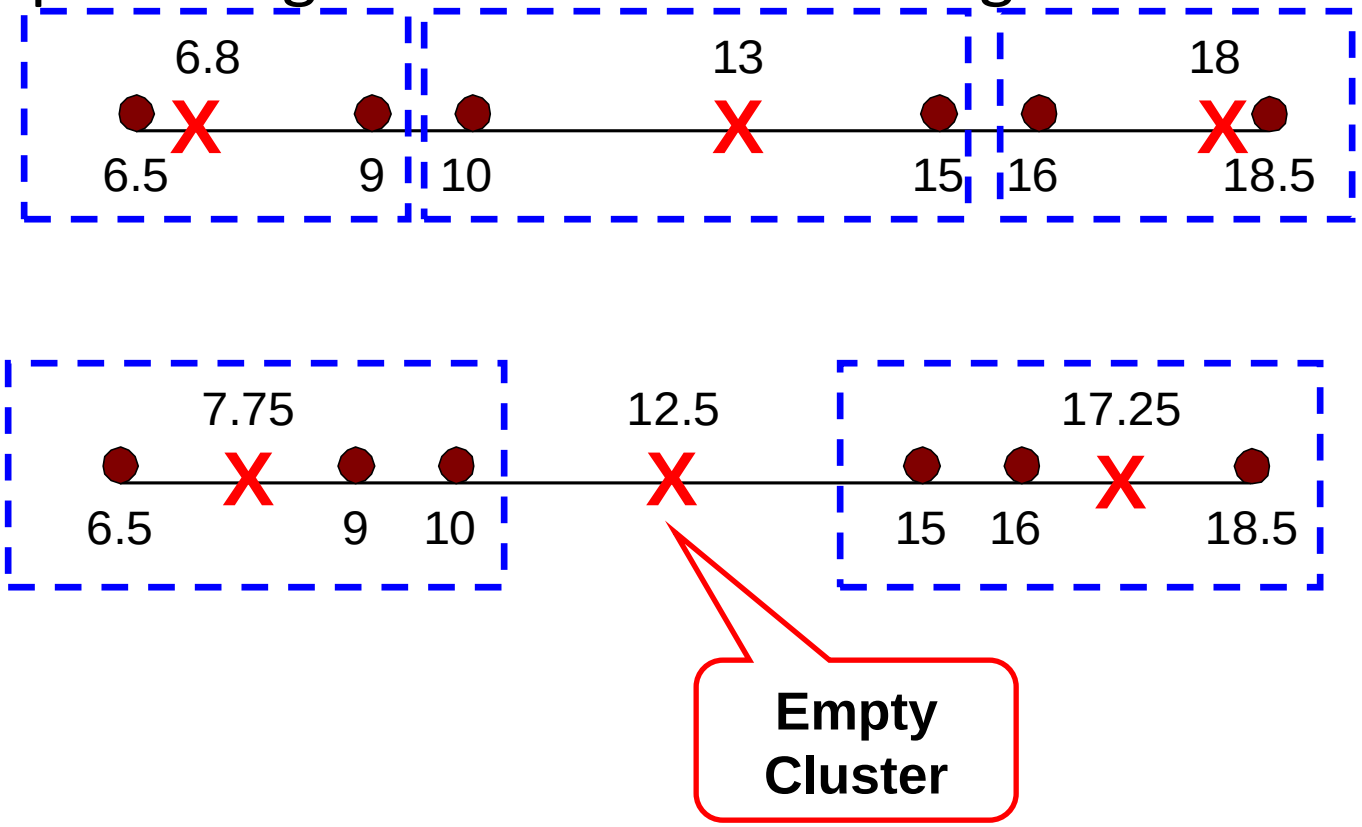
Pilih centroid baru secara acak dengan memilih titik dengan probabilitas

sebanding dengan $\frac{\min_j \text{d2(Cj, xi)}}{\sum_i \min_j \text{d2(Cj, xi)}}$ is

End For

Empty Clusters

- K-means dapat menghasilkan cluster kosong



Handling Empty Clusters

- Algoritma K-means dasar dapat menghasilkan kluster kosong

- Several strategies
  - Pilih poin yang paling berkontribusi pada SSE
  - Pilih titik dari klaster dengan SSE tertinggi
  - Jika ada beberapa cluster kosong, hal di atas dapat diulang beberapa kali.

Updating Centers Incrementally

- Dalam algoritma K-means dasar, centroid diperbarui setelah semua titik ditetapkan ke centroid

- Alternatifnya adalah memperbarui centroid setelah setiap penugasan (pendekatan inkremental)
  - Setiap tugas memperbarui nol atau dua centroid
  - Lebih mahal
  - Memperkenalkan dependensi pesanan
  - Jangan pernah mendapatkan cluster kosong
  - Dapat menggunakan "bobot" untuk mengubah dampak

Pre-processing and Post-processing

- Pre-processing
  - Menormalkan data
  - Hilangkan outlier
- Post-processing
  - Hilangkan kelompok kecil yang mungkin mewakili outlier
  - Memisahkan cluster 'longgar', yaitu cluster dengan SSE yang relatif tinggi
  - Gabungkan kluster yang 'dekat' dan memiliki SSE yang relatif rendah
  - Dapat menggunakan langkah-langkah ini selama proses pengelompokan
    - ISODATA

Bisecting K-means

- Membagi dua algoritma K-means
  - Varian K-means yang dapat menghasilkan pengelompokan partisi atau hierarkis

1: Initialize the list of clusters to contain the cluster containing all points.
2: **repeat**
3: Select a cluster from the list of clusters
4: **for** $i = 1$ to *number_of_iterations* **do**
5: Bisect the selected cluster using basic K-means
6: **end for**
7: Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: **until** Until the list of clusters contains K clusters

CLUTO: http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

Bisecting K-means Example