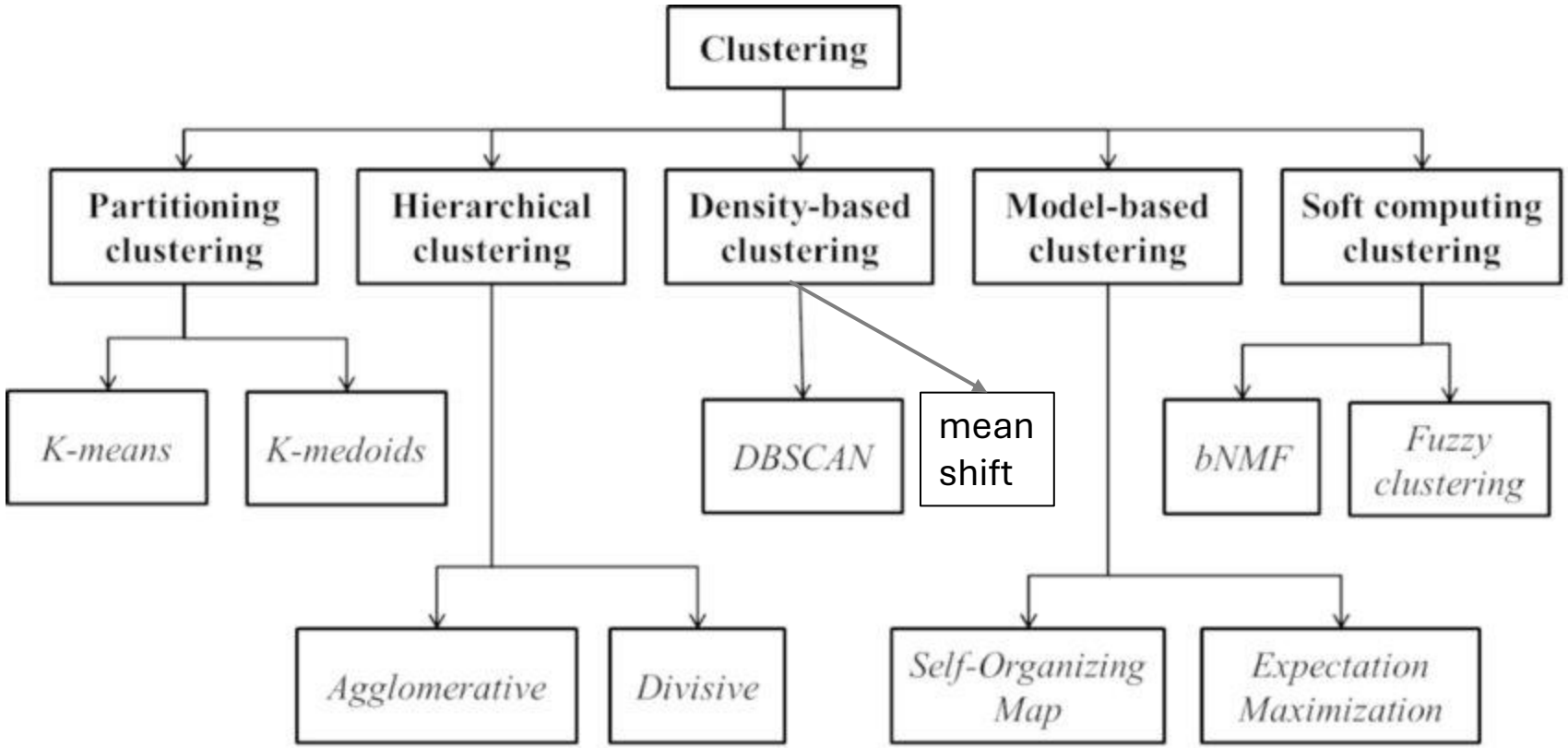
S1 Sains Data
FMIPA UNESA

Analisis Multivariat – Pertemuan 8

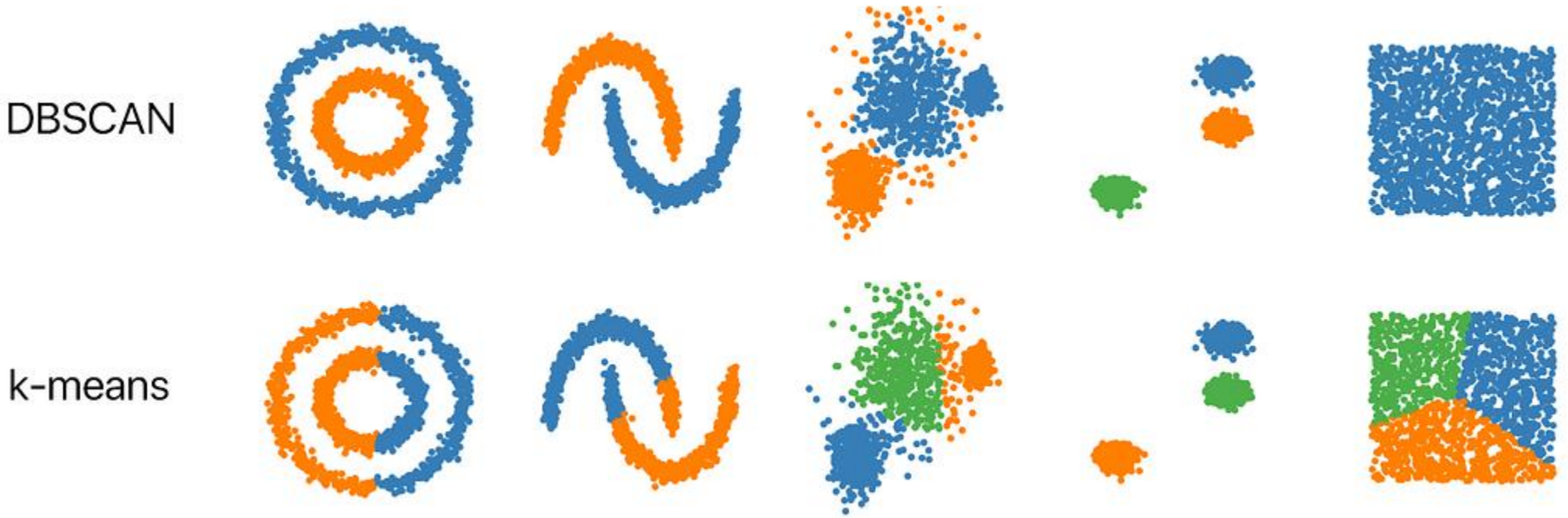
Clustering

Prodi S1 Sains Data
Universitas Negeri Surabaya
27 Maret 2025

DBSCAN


Clustering
Partitioning clustering
Hierarchical clustering
Density-based clustering
Model-based clustering
Soft computing clustering
K-means
K-medoids
DBSCAN
mean shift
bNMF
Fuzzy clustering
Agglomerative
Divisive
Self-Organizing Map
Expectation Maximization

DBSCAN

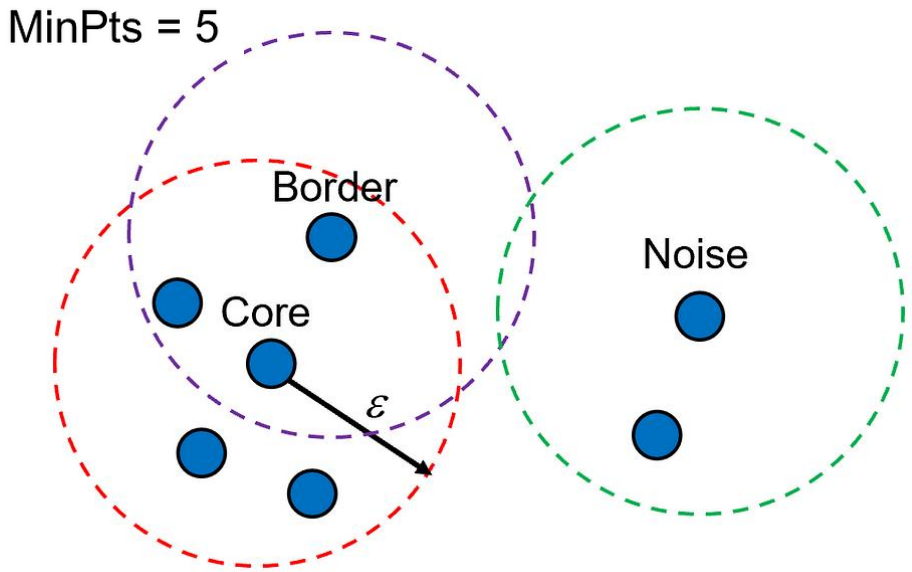


DBSCAN

- *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN)
- DBSCAN merupakan salah satu metode dari density-based clustering.
- Density-based clustering
  - digunakan untuk memperoleh atau menghasilkan pola yang sebelumnya tidak diketahui dari sekumpulan data.
  - menempatkan daerah dengan kepadatan tinggi dan dipisahkan satu dengan yang lain oleh daerah dengan kepadatan rendah.
- Pada DBSCAN, density (kepadatan) dihitung dengan menghitung jumlah objek dalam lingkungan yang ditentukan oleh parameter radius (Eps).

Konsep DBSCAN

- **Core points** adalah titik yang berada di dalam wilayah yang padat.
  - Titik yang memiliki cukup banyak tetangga dalam radius ε (minimal MinPts).
  - Titik-titik ini adalah pusat utama dalam pembentukan cluster.
- **Border points**
  - Bukanlah core points namun berada di lingkungan core points atau titik yang berada di tepi wilayah yang padat.
  - Jumlah titik di sekitar border points masih berada dalam radius yang ditentukan (Eps), kurang dari batas minimal yang telah ditetapkan (MinPts).
- **Noise points** berada di wilayah yang jarang ditempati.



Parameter DBSCAN

- **ε (epsilon) atau radius**: Jarak maksimum di sekitar titik yang menentukan lingkungan (neighborhood) titik tersebut.
- **MinPts (minimum objek)**: Jumlah minimal objek yang harus ada dalam lingkungan suatu titik agar titik tersebut dapat membentuk cluster.

$$d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

defining 3 terms, required for understanding the DBSCAN algorithm

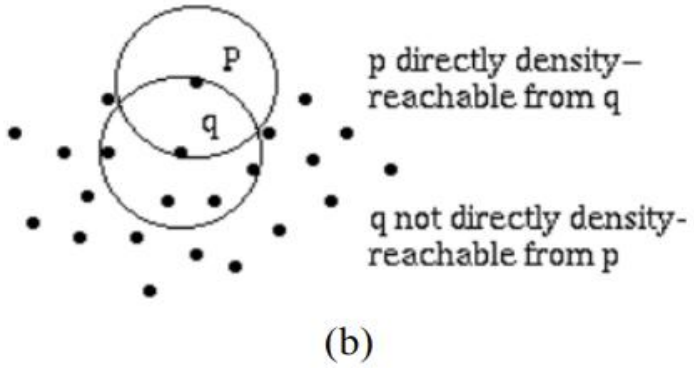


Fig. 3. (b) Directly-density-reachable [11]

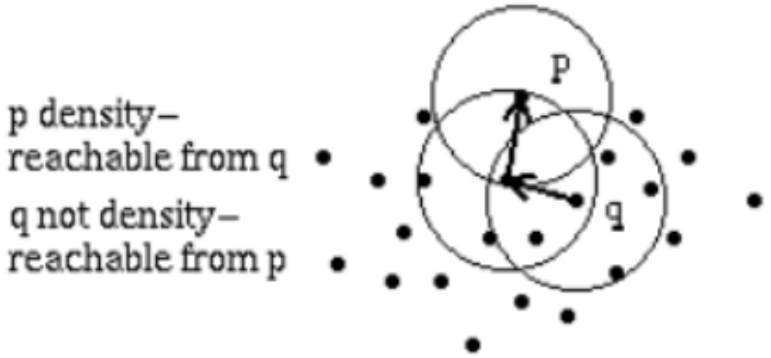


Fig. 4. Density-reachable [11]

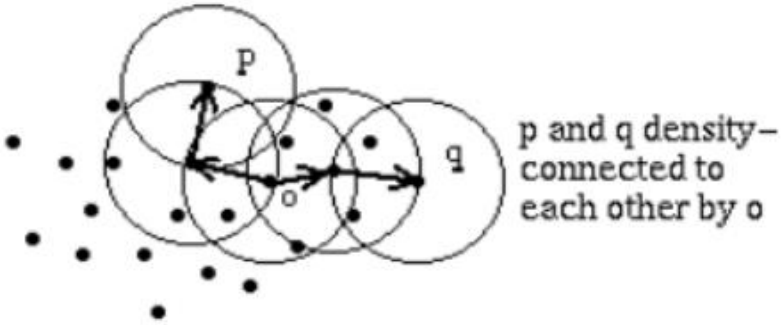


Fig. 5. Density-connected [11]

A point is called direct density reachable if it has a core point in its neighbourhood.

semua titik dalam radius **ε** dari **q** adalah **direct density reachable** dari **q**.

A point is called density reachable from another point if they are connected through a series of core points.

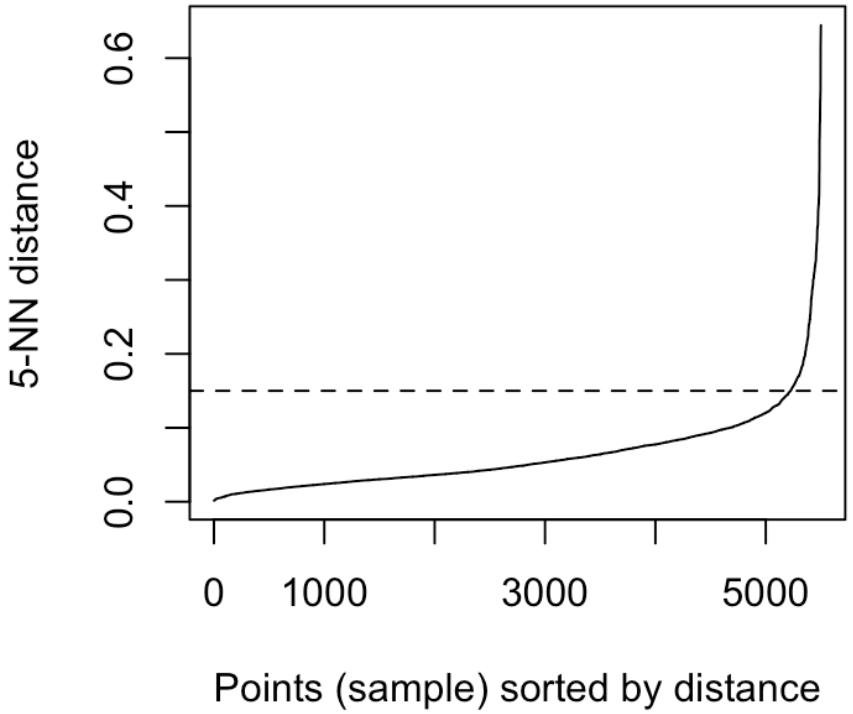
Two points are called density connected if there is a core point which is density reachable from both the points.

Langkah-langkah

1. Inisialisasi parameter MinPts dan Eps.
2. Pilih satu titik *p* yang belum diklasifikasikan.
3. Menghitung jumlah tetangga dalam lingkungan titik *p* yang ditentukan oleh parameter radius (Eps).
   - Apabila jumlah titik di sekitar titik p kurang dari MinPts, maka titik p diklasifikasikan sebagai border point dan kembali ke langkah 2.
   - Sementara itu, apabila jumlah titik di sekitar titik p lebih dari atau sama dengan MinPts, maka titik p diklasifikasikan sebagai core point dan dibentuk cluster baru serta lanjut ke langkah 4.
4. Memasukkan semua titik dalam lingkungan core point ke dalam cluster.
5. Melakukan identifikasi pada titik-titik yang berada dalam lingkungan core point.
   - Apabila ditemukan core point baru maka kembali ke langkah 4.
   - Sementara itu, apabila tidak ditemukan lagi core point dalam lingkungan maka pembentukan cluster yang bersangkutan berhenti.
6. Mengulangi langkah 2 sampai 5 hingga semua titik telah diklasifikasikan.
7. Titik yang tidak masuk ke dalam cluster manapun dianggap sebagai noise points.

Menentukan Nilai ε (Epsilon)

1. **Domain knowledge**
2. **Grafik K-Distance**
   - Hitung jarak k-tetangga (biasanya k = minPts).
   - Buat plot k-Nearest Neighbor Distance, yaitu sumbu X adalah titik data, dan sumbu Y adalah jarak ke tetangga ke-k.
   - Cari elbow point: Titik di mana jarak mulai meningkat tajam menunjukkan nilai ε yang optimal.



Menentukan MinPts

- minPts≥D+1
- The larger the data set, the larger the value of minPts should be chosen. minPts must be chosen at least 3.

Kelebihan

- Mampu mengidentifikasi cluster dengan bentuk arbitrer (tidak hanya bulat atau persegi).
- Tidak membatasi jumlah objek dalam setiap cluster.
- Dapat menangani noise dalam data (objek yang berada di area sangat jarang).
- Tidak perlu menentukan jumlah cluster terlebih dahulu seperti pada K-Means.

Kekurangan

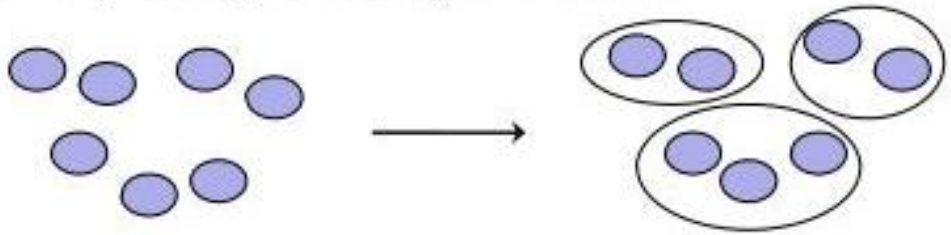
- Sulit menentukan nilai ε yang optimal.
- Tidak bekerja dengan baik pada dataset dengan kepadatan yang sangat bervariasi.
- Sensitif terhadap pemilihan parameter MinPts dan ε.

FUZZY C-MEANS

Fuzzy C-Means

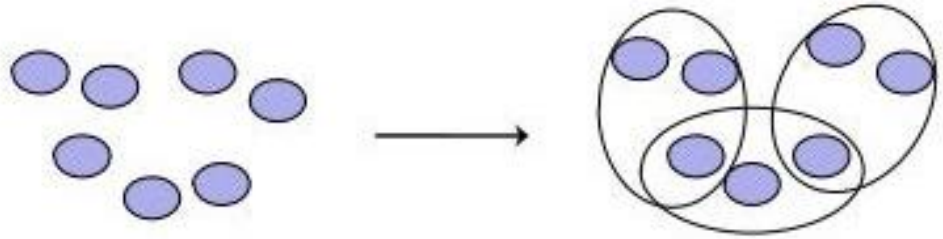
Hard Clustering

- Every object may belong to exactly one cluster.



Soft Clustering

- The membership is fuzzy - Objects may belong to several clusters with a fractional degree of membership in each.



Fuzzy C-Means

- Metode Fuzzy C-Means (FCM) merupakan pengembangan dari K-Means dengan menerapkan sifat fuzzy keanggotaannya.
- Dalam FCM, setiap titik data memiliki derajat keanggotaan (membership function, u_{ik}) terhadap setiap cluster.
  - Menunjukkan seberapa besar kemungkinan suatu data bisa menjadi anggota ke dalam suatu kelompok
  - Derajat ini bernilai antara 0 dan 1 dan menunjukkan seberapa kuat data tersebut terkait dengan suatu cluster.
- Didasarkan pada fungsi objektif yang **meminimalkan jarak berbobot antara titik data dan pusat cluster**, dengan mempertimbangkan derajat keanggotaan fuzzy.
- FCM memperkenalkan suatu variabel m yang merupakan weighting exponent dari membership function. Variabel ini dapat mengubah besar pengaruh dari membership function, dalam proses clustering menggunakan metode FCM, m mempunyai wilayah nilai lebih besar dari 1 (m>1).

Langkah-Langkah

1. Inisialisasi jumlah cluster C, parameter fuzziness m (biasanya m>1), dan pusat cluster awal secara acak.
2. Hitung derajat keanggotaan, uij
3. Perbarui pusat cluster Cj
4. Hitung jarak setiap objek ke pusat klaster
5. Cek konvergensi. Jika perubahan derajat keanggotaan atau pusat cluster kecil (di bawah threshold tertentu), hentikan iterasi. Jika tidak, ulangi dari langkah 2.

Membership Degree

$$u_{ik} = \sum_{i=1}^{c} \left[\left(\frac{D(x_k, v_i}{D(x_k, v_j)} \right)^{\frac{2}{m-1}} \right]^{-1}$$

di mana:

u_{ik} = *Membership function* data ke-*k* ke kelompok ke-*i*
v_i = Nilai *centroid* kelompok ke-*i*
v_j = Nilai *centroid* kelompok ke-*j*
m = *Weighting exponent*
c = Banyaknya *cluster*

- Membership function mempunyai jangkauan nilai $0 \leq u_{ik} \leq 1$.
- Data item yang mempunyai tingkat kemungkinan yang lebih tinggi ke suatu kelompok akan mempunyai nilai membership function ke kelompok tersebut yang mendekati angka 1 dan kelompok yang lain mendekati angka 0, dengan syarat $1 < m < \infty$, $0 \leq u_{ik} \leq 1$, $\sum_{i=1}^{c} u_{ik} = 1$

Update Centroid

Nilai *centroid cluster*

$$v_{i=}\frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}}$$

n = Banyaknya data
k = Variabel ke-k
i = Kelompok ke-
c = Banyaknya *cluster*
v_i = Niali *centroid* kelompok ke-i
m = *Weighting exponent*

Kelebihan

- Lebih fleksibel karena memungkinkan data masuk ke beberapa cluster.
- Lebih akurat untuk data dengan transisi antar cluster yang tidak jelas.
- Bisa menangani data yang tumpang tindih lebih baik dibanding K-Means.

Kekurangan

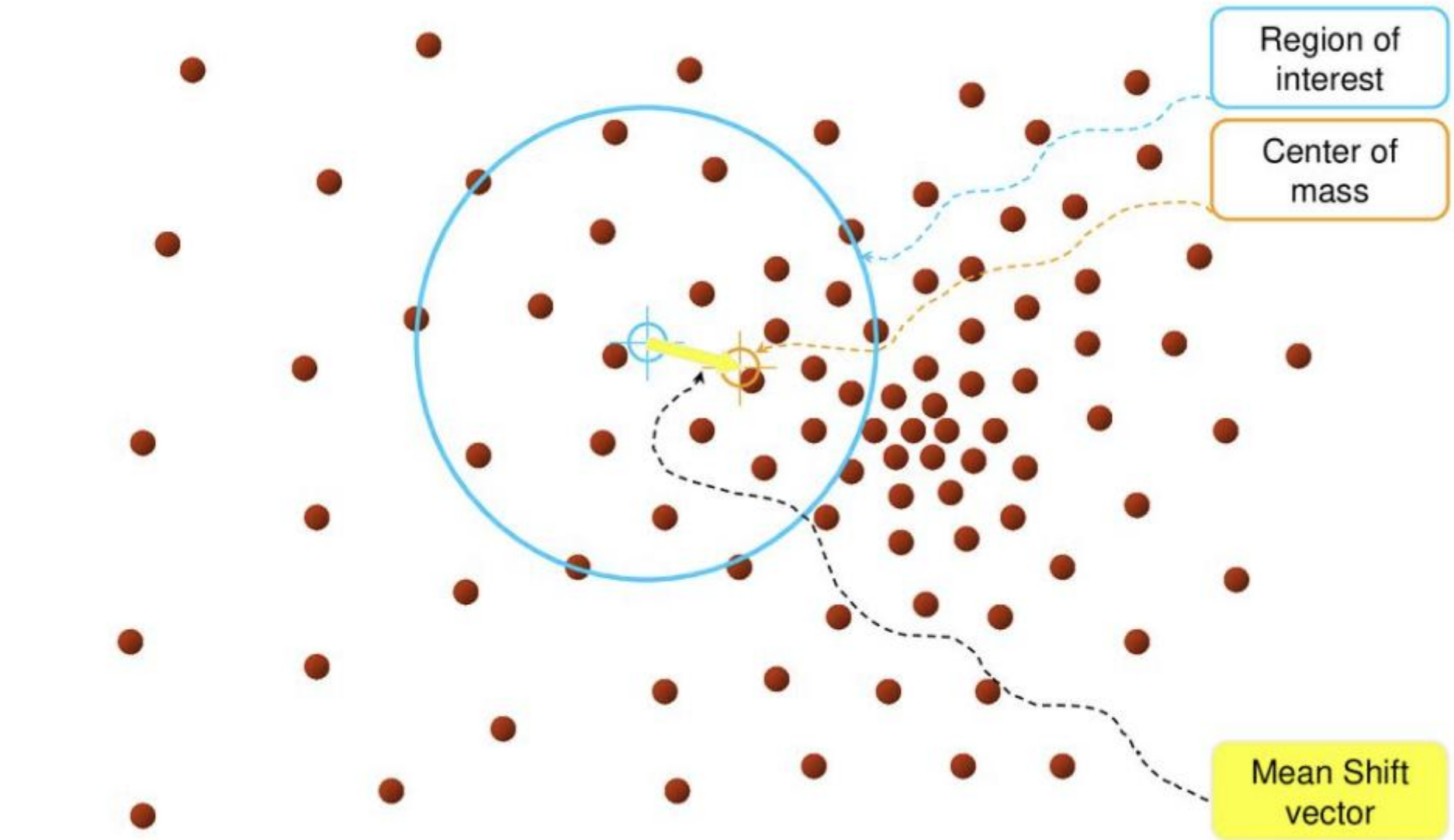
- Lebih komputasi-intensif dibanding K-Means.
- Harus menentukan nilai fuzziness m yang optimal.
- Sensitif terhadap outlier.

Penerapan FCM

- **Segmentasi citra** (misalnya dalam pengolahan gambar medis).
- **Analisis pasar** (menentukan segmentasi pelanggan).

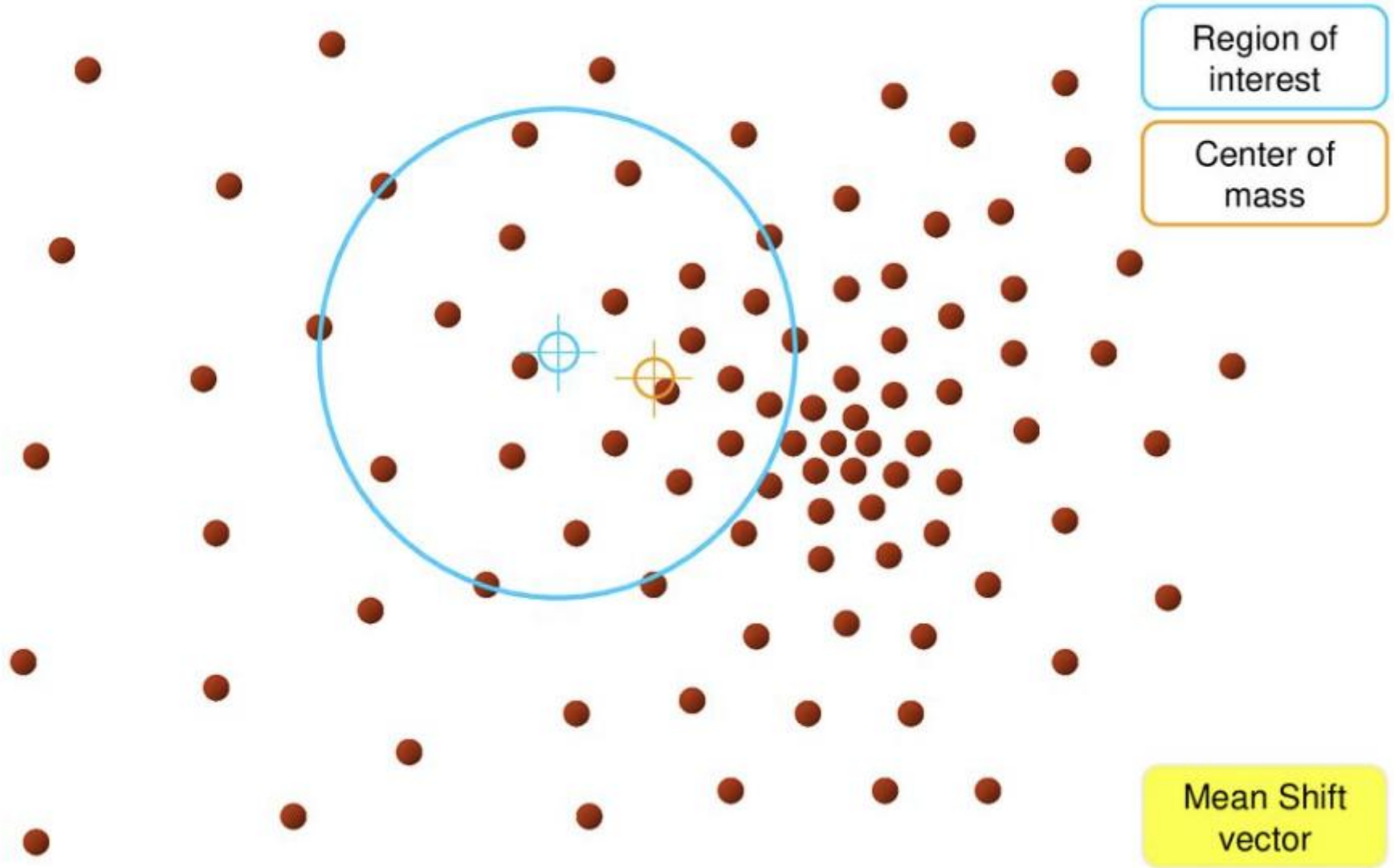
MEAN SHIFT

Mean-Shift



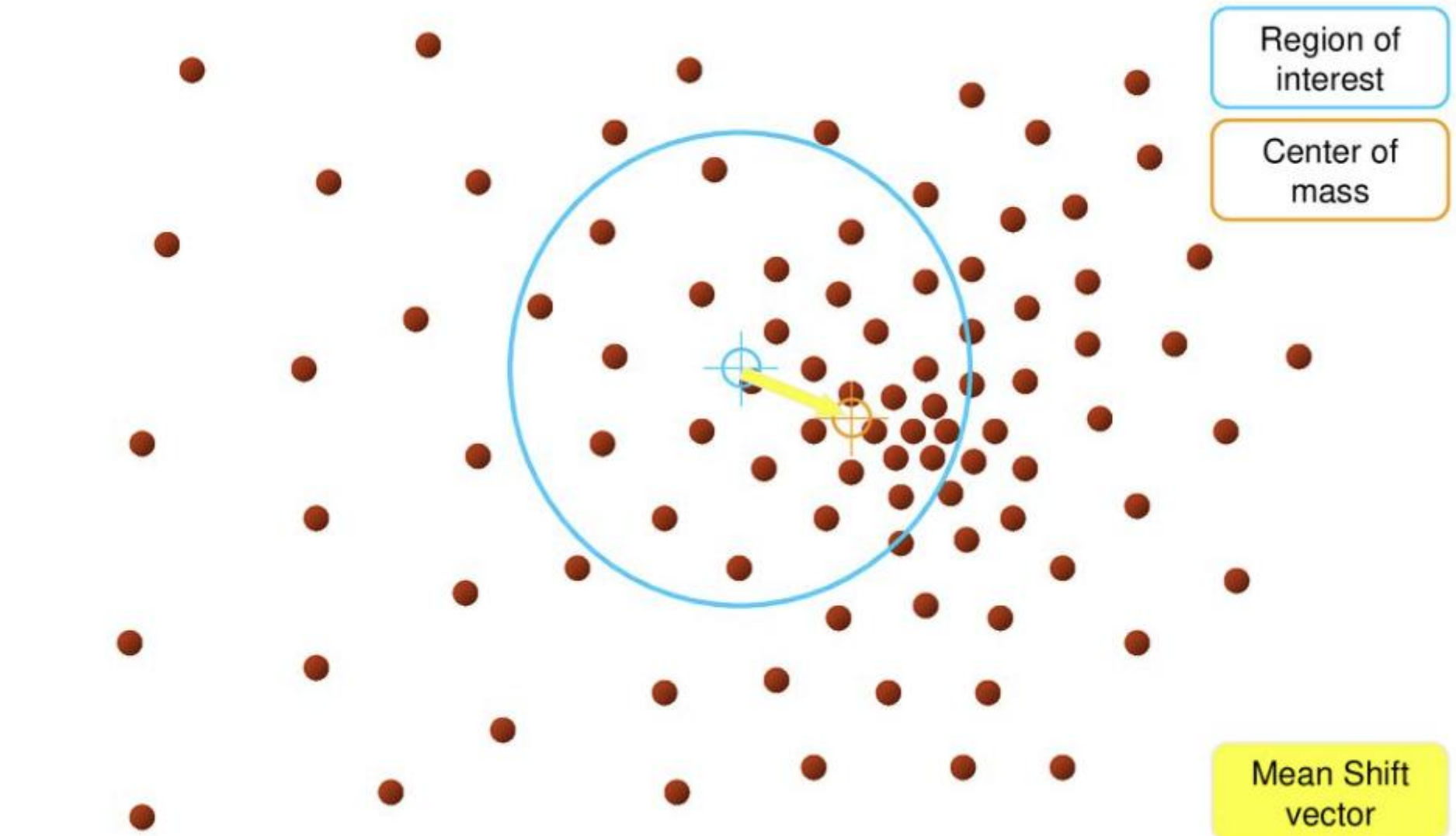
Slide by Y. Ukrainitz & B. Sarel

Mean-Shift



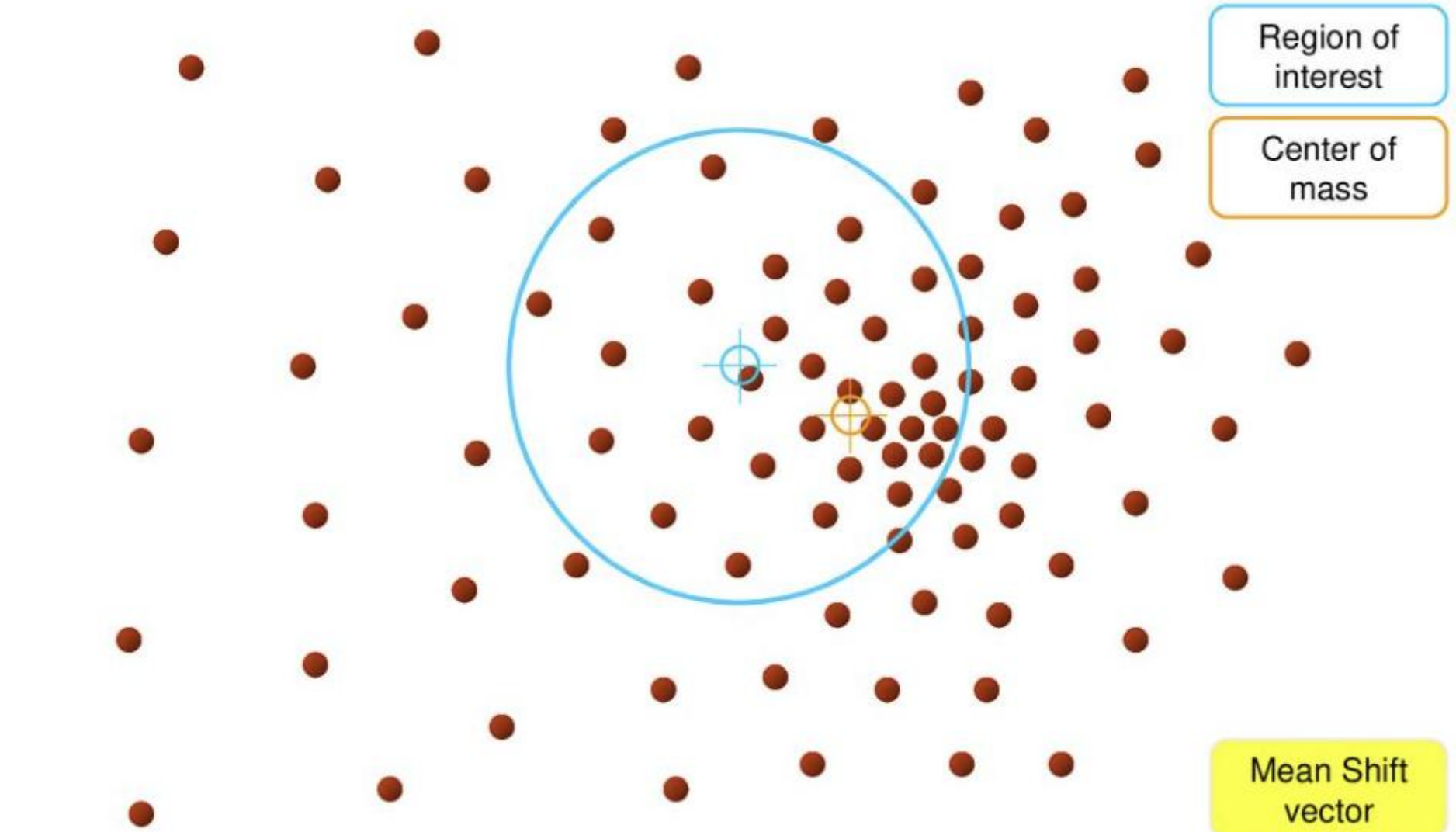
Slide by Y. Ukrainitz & B. Sarel

Mean-Shift



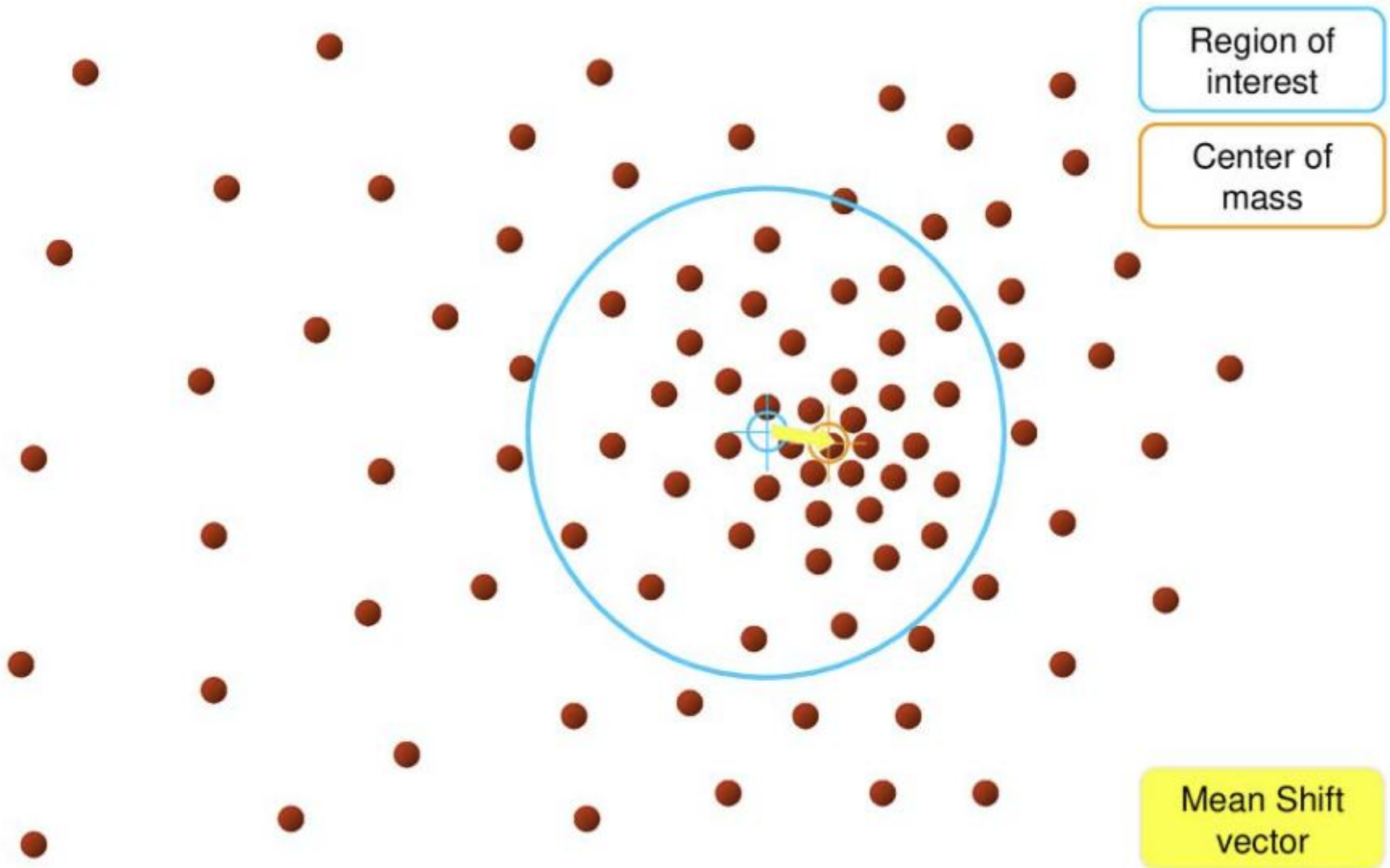
Slide by Y. Ukrainitz & B. Sarel

Mean-Shift



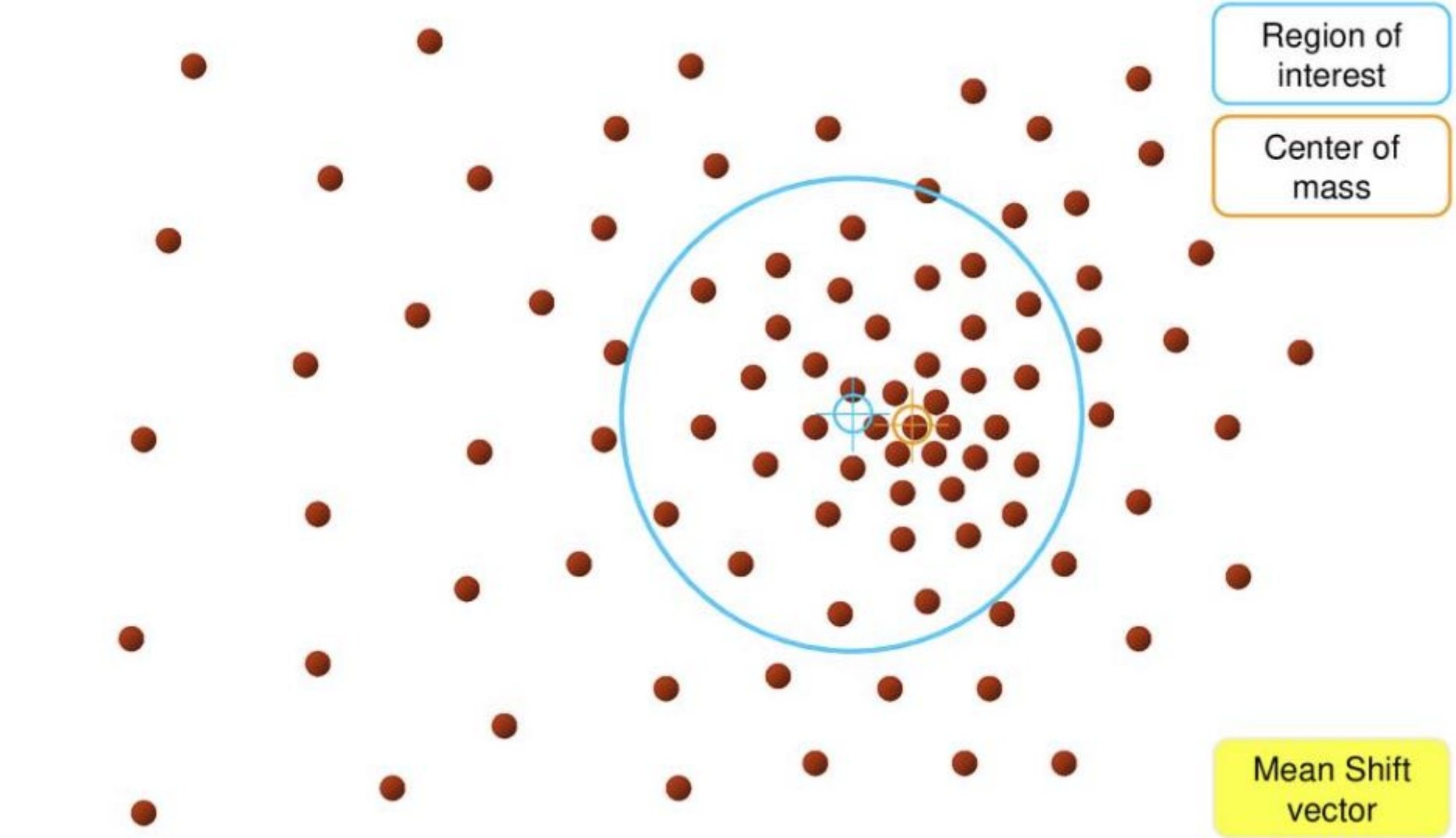
Slide by Y. Ukrainitz & B. Sarel

Mean-Shift



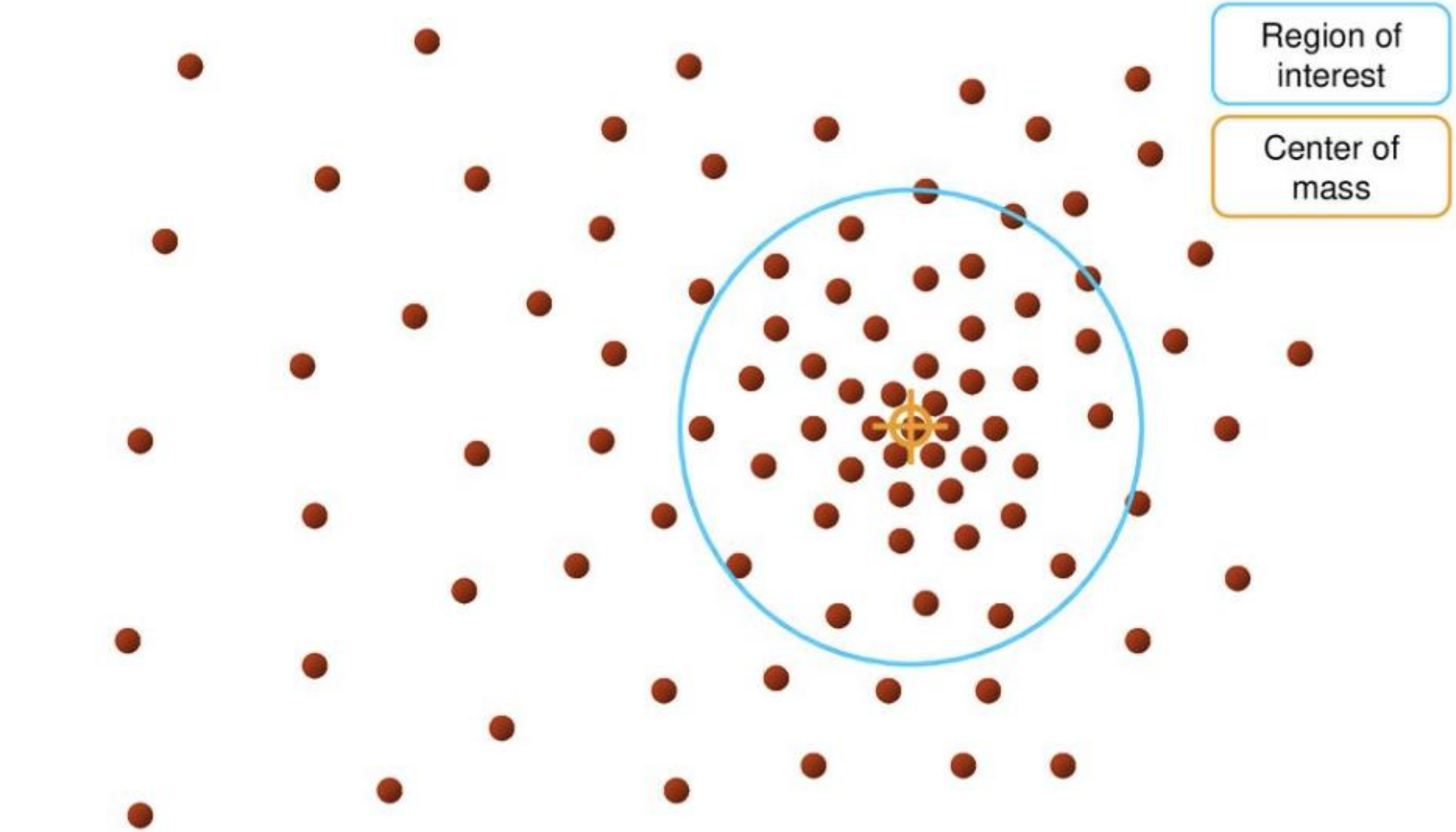
Slide by Y. Ukrainitz & B. Sarel

Mean-Shift



Slide by Y. Ukrainitz & B. Sarel

Mean-Shift



Slide by Y. Ukrainitz & B. Sarel

Mean Shift Clustering

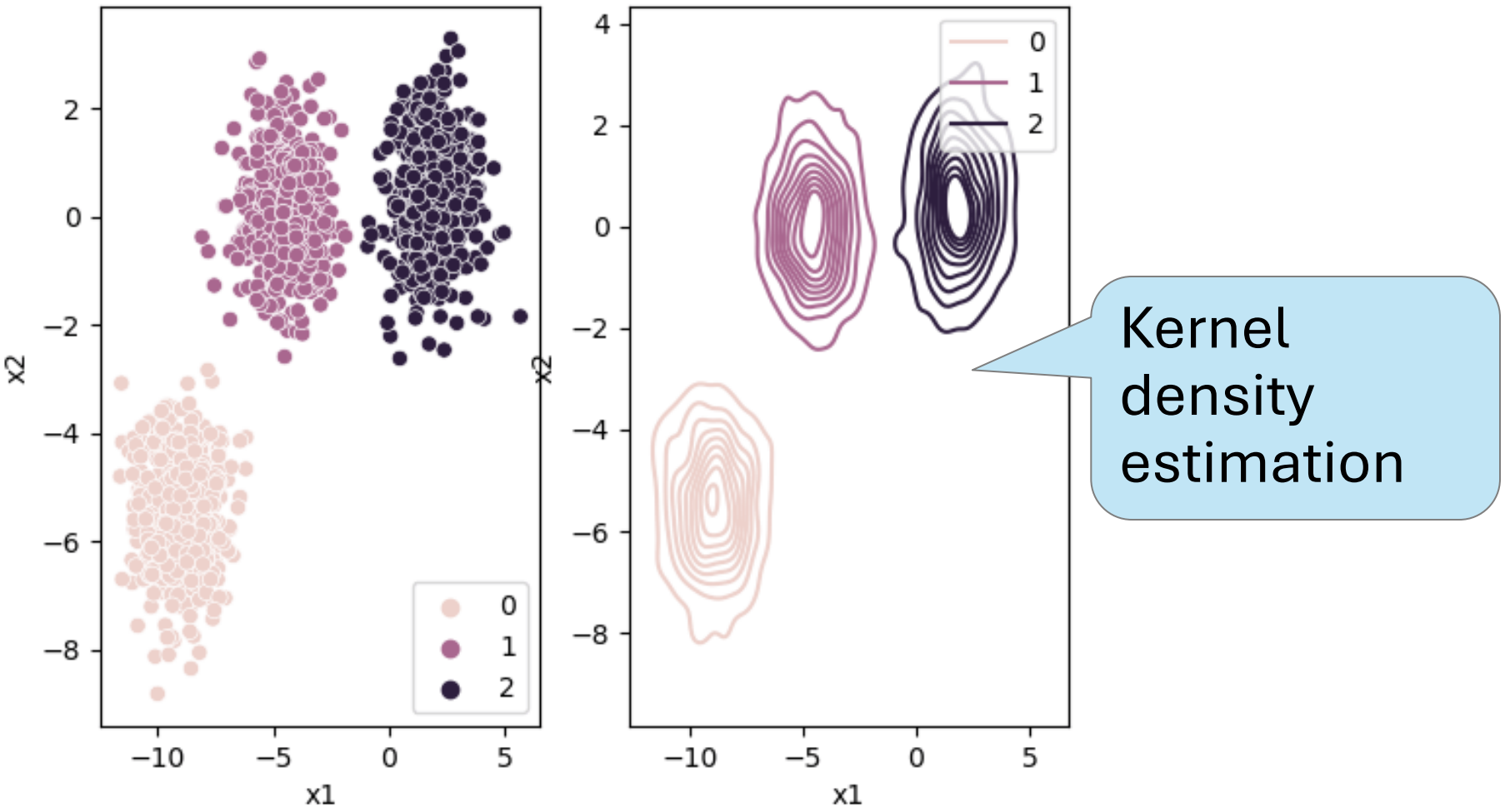
- Metode clustering berbasis densitas yang tidak memerlukan jumlah cluster sebagai parameter awal.
- Algoritma ini bekerja dengan **mencari daerah dengan kepadatan data tinggi (modes)** melalui pendekatan berbasis estimasi kernel density estimation (KDE).
  - Titik-titik data akan berpindah ke pusat distribusi data di sekitarnya,

Langkah-langkah

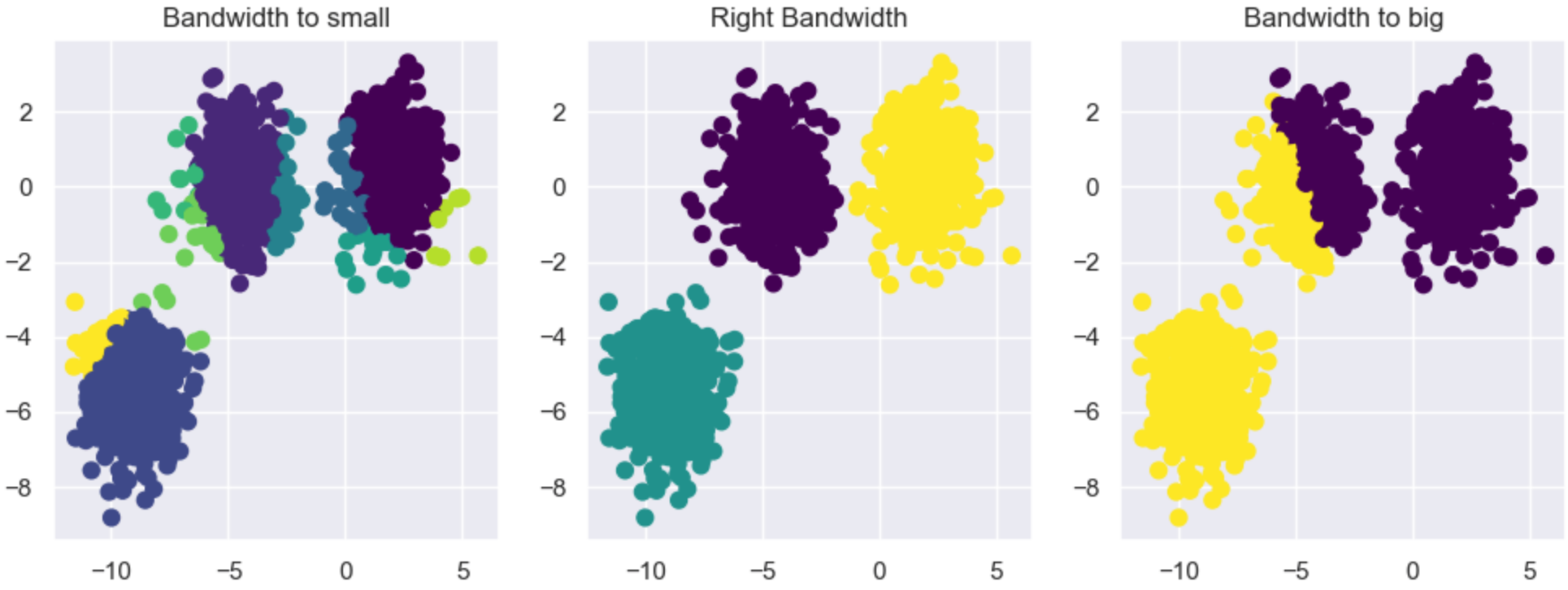
1. Tentukan kernel dan bandwidth (h)
2. Pilih titik data awal sebagai centroid awal.
3. Hitung mean shift vector:
   - Untuk setiap titik xi, cari semua titik lain dalam radius bandwidth h.
   - Hitung mean center dari titik-titik tersebut:

$$X_{\text{new}} = \frac{\sum_j K(d) \cdot X_j}{\sum_j K(d)} \qquad Y_{\text{new}} = \frac{\sum_j K(d) \cdot Y_j}{\sum_j K(d)}$$

   - Geser titik xi menuju mean center m.
4. Ulangi langkah 3 hingga konvergen (pergerakan titik sangat kecil).
5. Kelompokkan titik-titik yang mendekati mode yang sama menjadi satu cluster


x2
x1
0
1
2
Kernel density estimation

Memilih Bandwidth (*h*)



- Gaussian Kernel

Most often, we require the covariance matrix being identity, hence:

$$K(d) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{d}{h}\right)^2\right)$$

- Uniform Kernel

$$K(x) = 1\{\|x\| < 1\}.$$

- Epanechnikov Kernel

$$K(\boldsymbol{x}) = \frac{3}{4}(1 - |x|^2)1\{|x| < 1\}.$$

Kelebihan

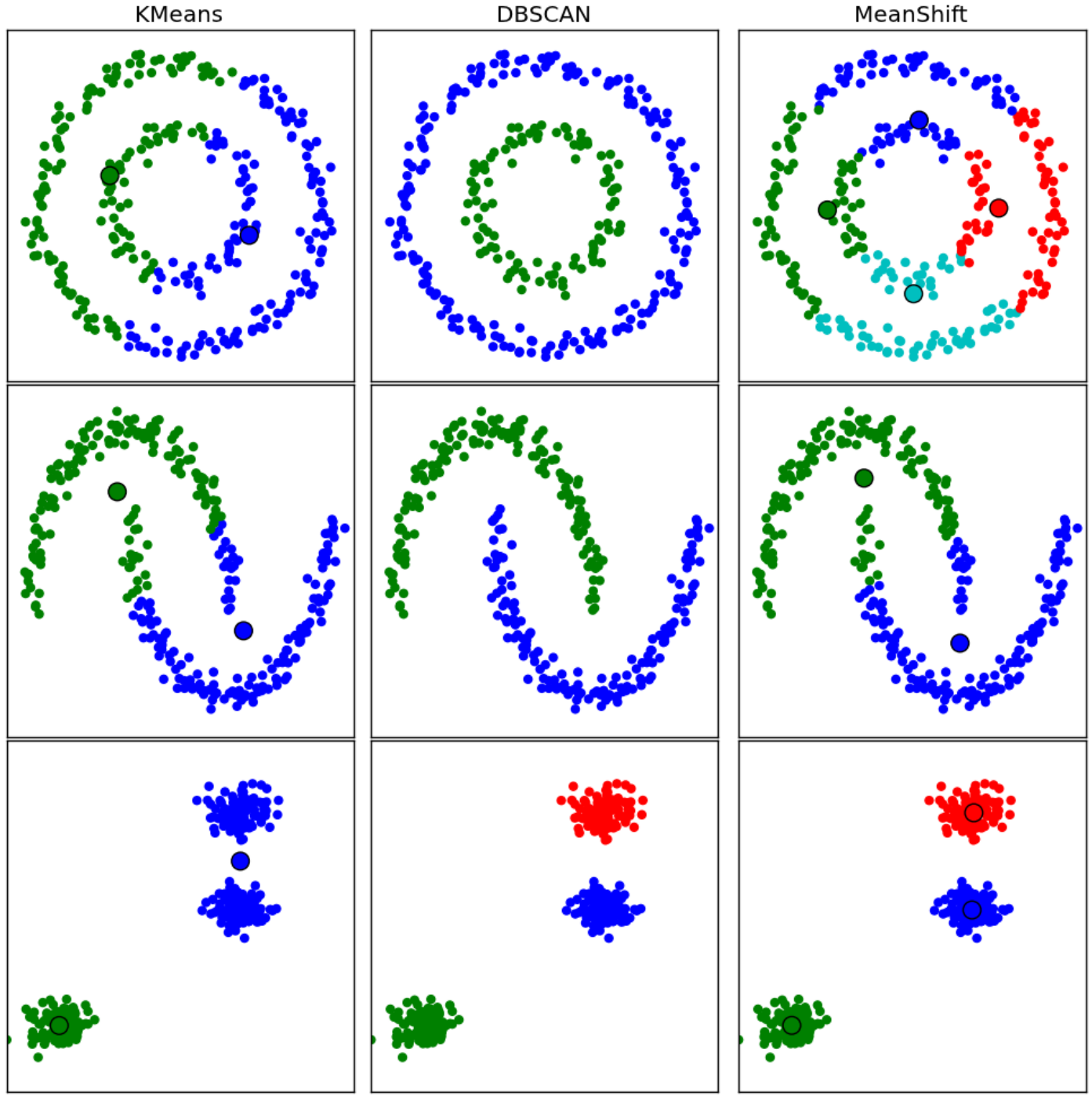
- Tidak perlu menentukan jumlah cluster sebelumnya.
- Mampu menangani cluster dengan bentuk tidak teratur.
- Secara alami menangani outlier tanpa perlu parameter tambahan.

Kekurangan

- Pemilihan bandwidth h sangat mempengaruhi hasil.
- Komputasi mahal, terutama jika jumlah data besar.
- Tidak selalu efisien untuk data berdimensi tinggi.

Penerapan Mean Shift

- Segmentasi gambar
- Pelacakan objek dalam video.
- Analisis data pasar untuk menemukan kelompok pelanggan potensial.
- Deteksi anomali dalam data keuangan.


KMeans
DBSCAN
MeanShift

Referensi

Nisa, K. K., Andrianto, H. A., & Mardhiyyah, R. (2014). *Hotspot clustering using DBSCAN algorithm and shiny web framework. 2014 International Conference on Advanced Computer Science and Information System*.

Manual Clustering

https://docs.google.com/spreadsheets/d/1icvoo2R6kEi3DYtYN38CanK-ezZRriX1/edit?usp=sharing&ouid=103490478088003213054&rtpof=true&sd=true