

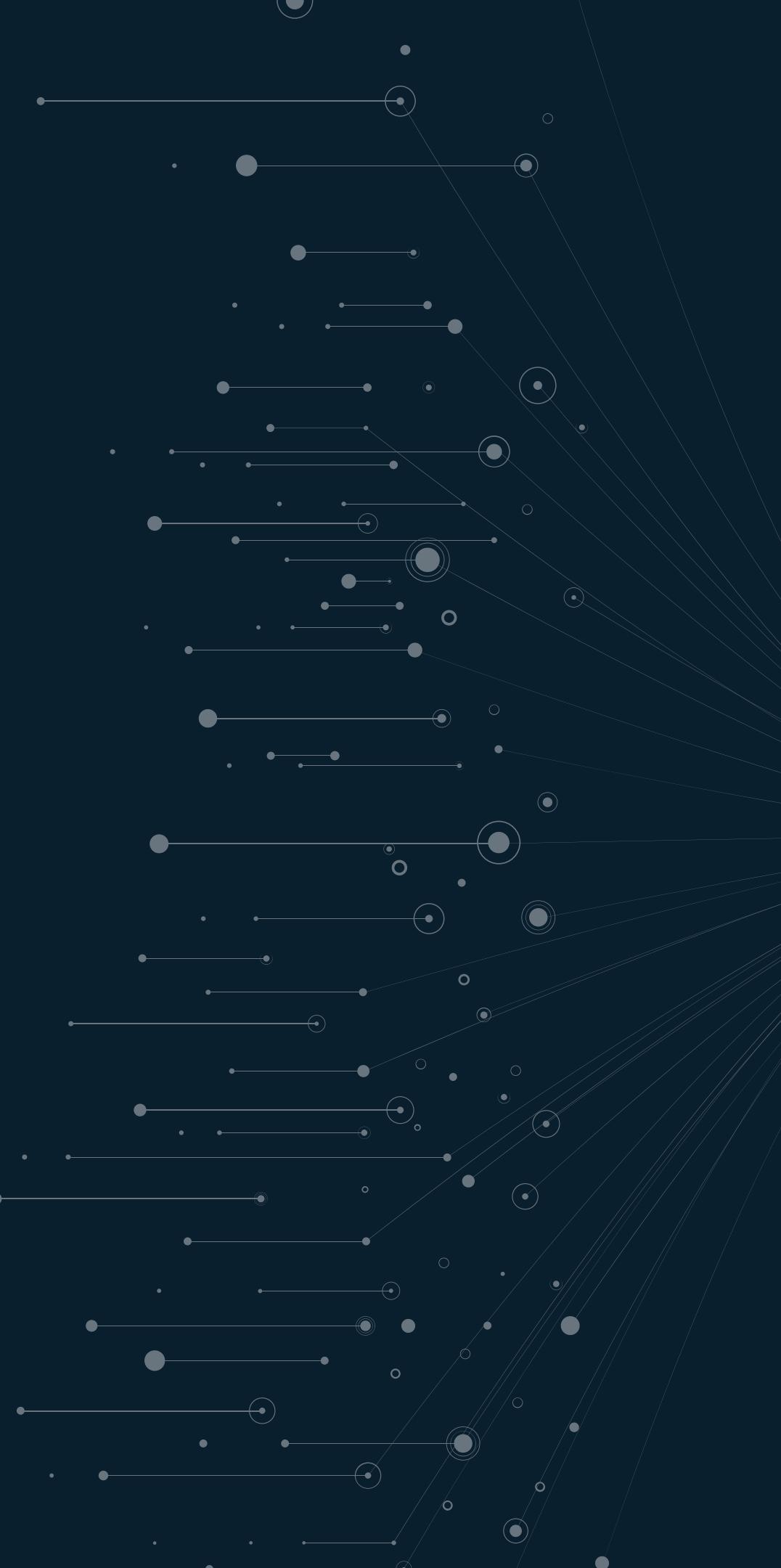
Đồ Án KTHP Môn học Máy Học

# ÁP DỤNG THUẬT TOÁN MEAN SHIFT CHO BỘ DỮ LIỆU SHOP CUSTOMER DATA

GVHD: TS. Nguyễn An Tế  
Nhóm: 10

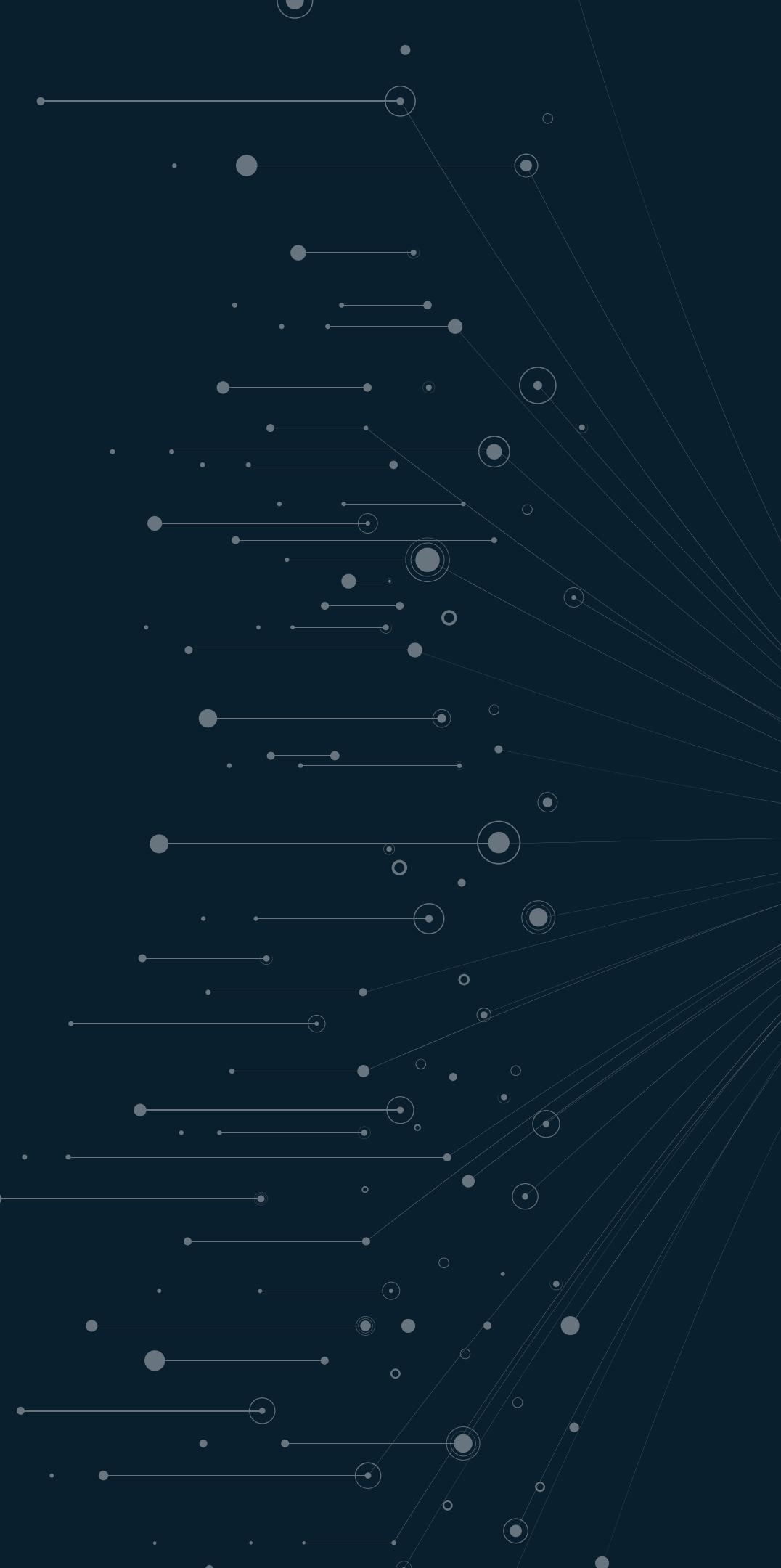
# THÀNH VIÊN

1. Nguyễn Thị Tuyết
2. Vũ Nguyễn Thảo Vi
3. Bùi Quốc Việt
4. Nguyễn Quốc Việt
5. Nguyễn Nhật Thảo Vy



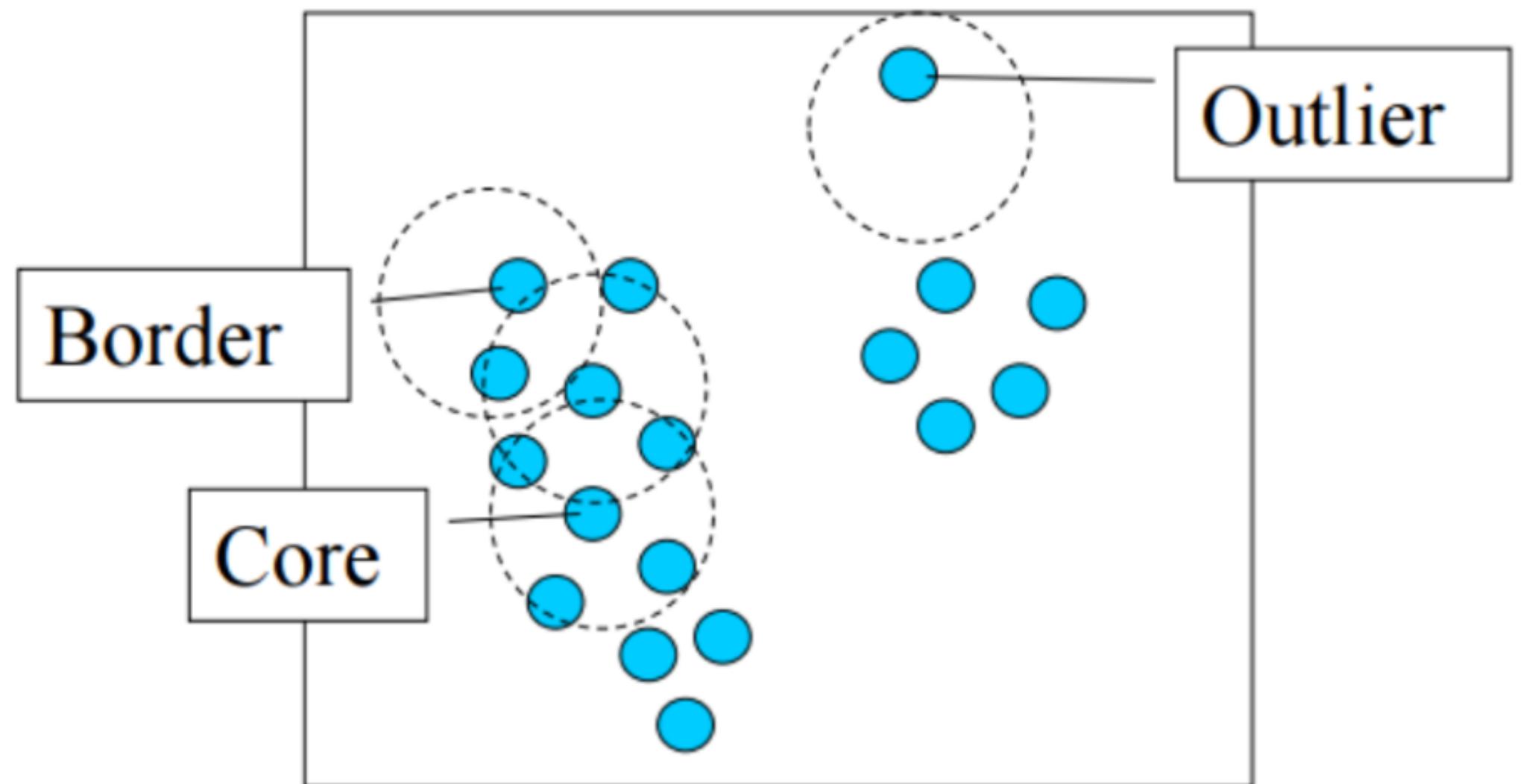
Đồ Án KTHP Môn học Máy Học

# THUẬT TOÁN MEAN SHIFT

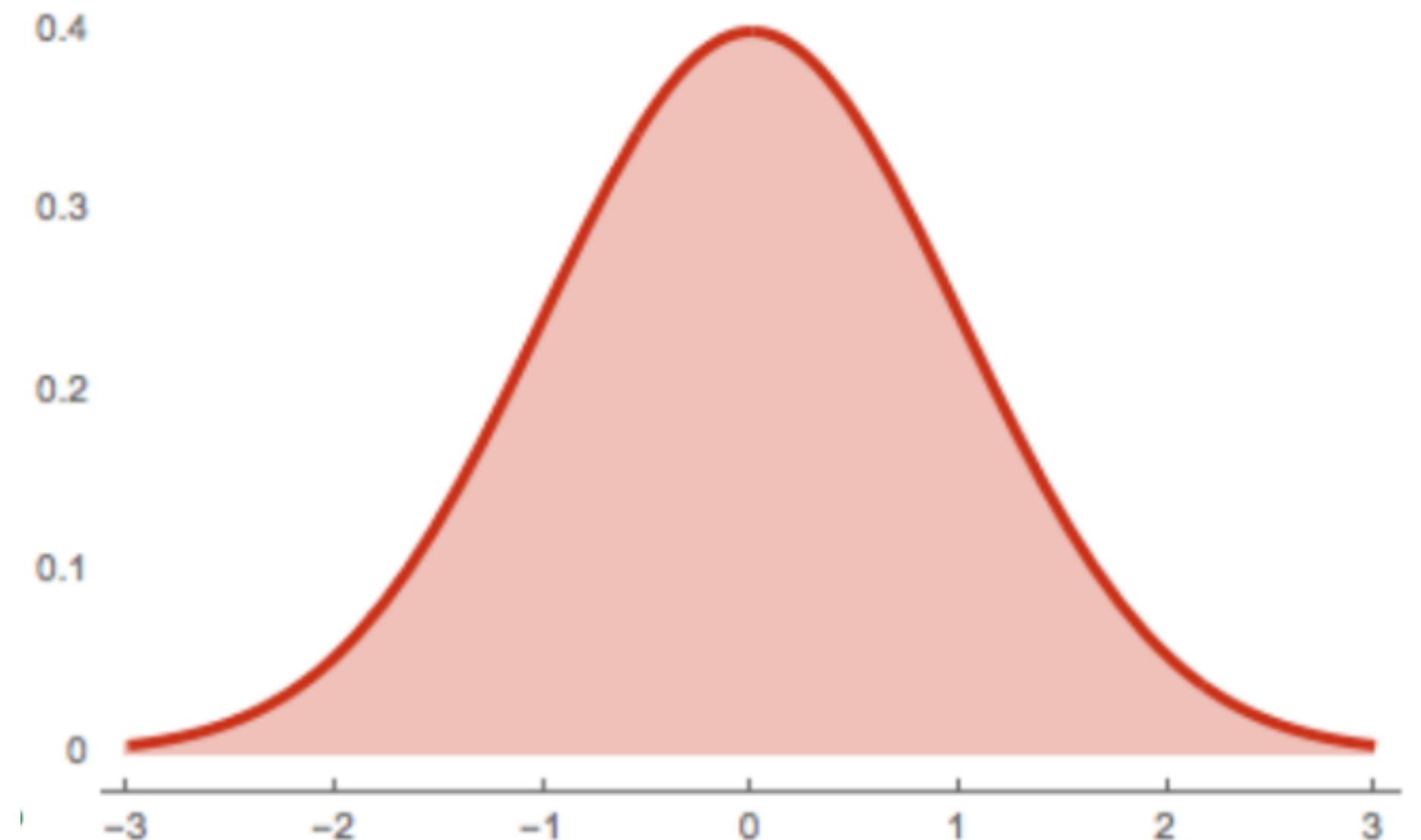


# Density-Based Clustering method

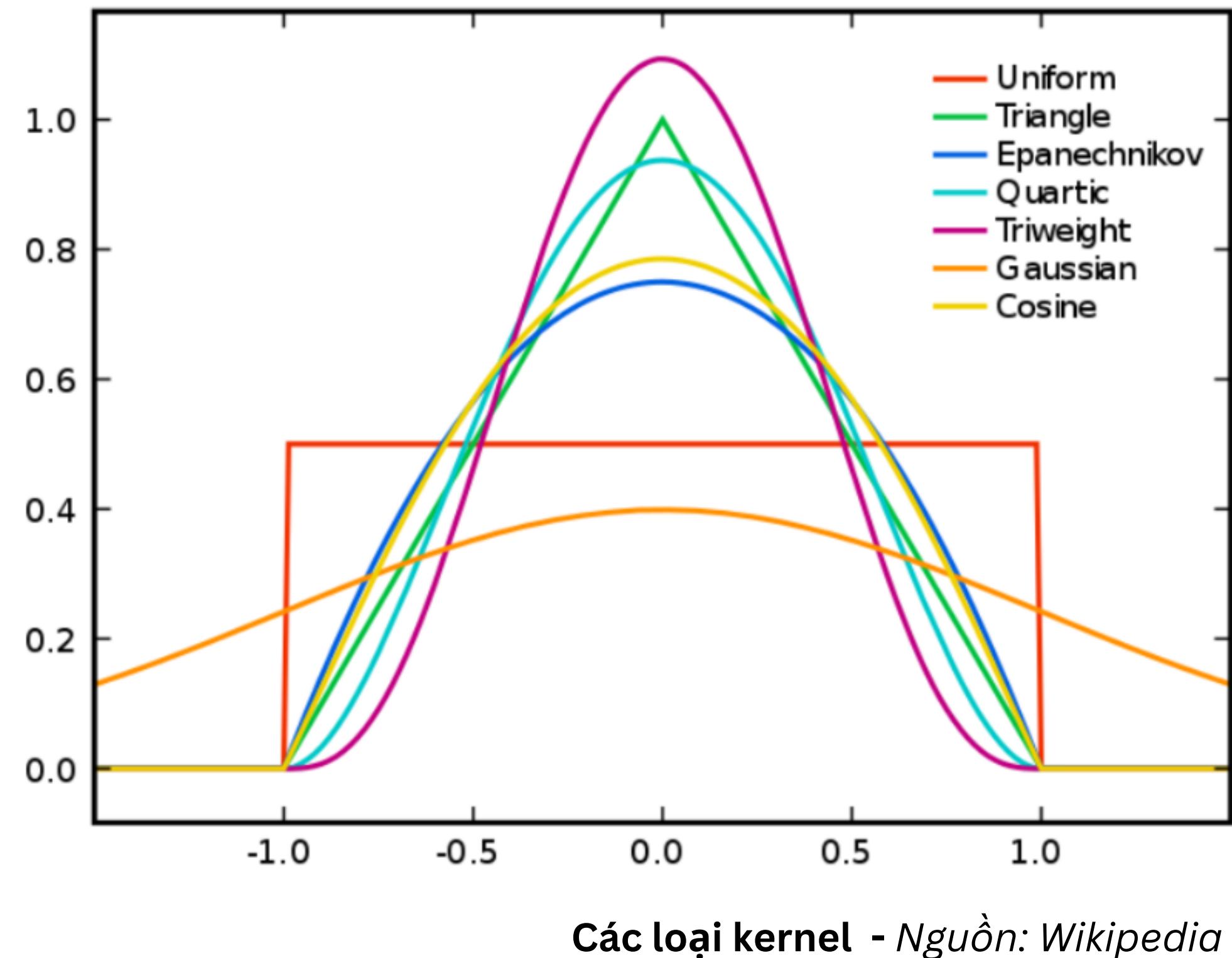
Mean Shift là một  
thuật toán phân cụm  
dựa trên mật độ.



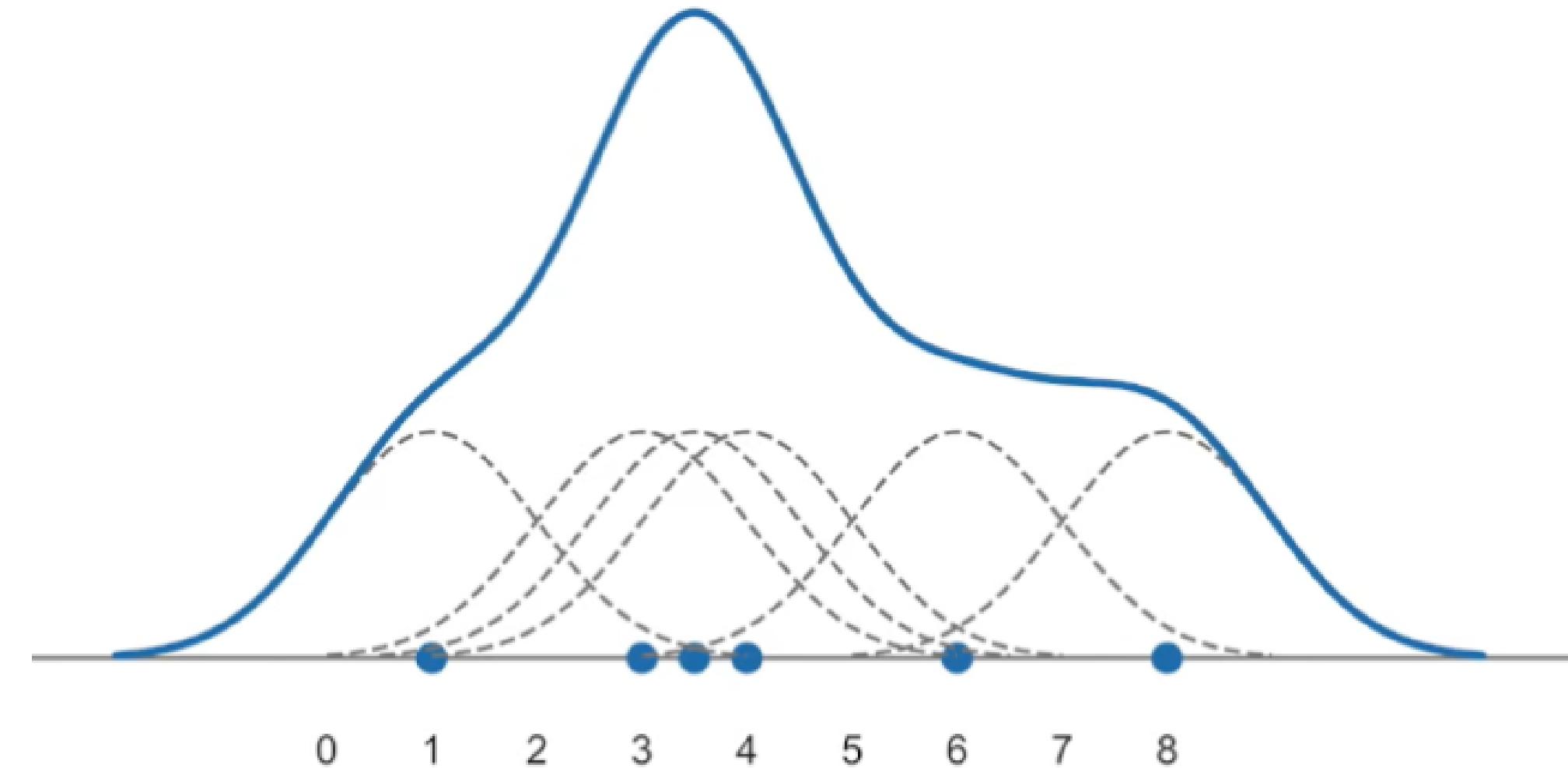
# Probability Density Function - PDF



# Kernel Density Estimate - KDE

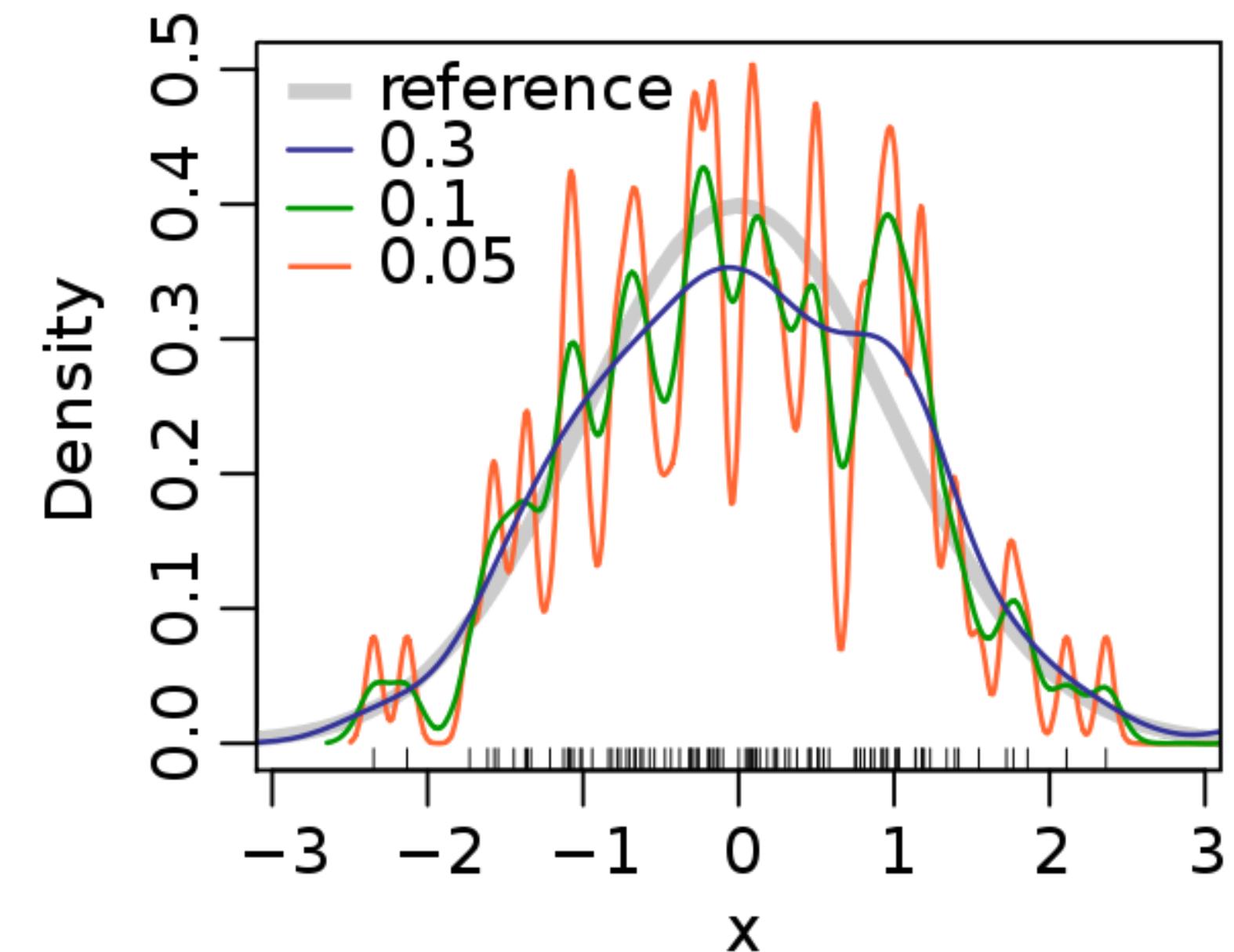


## Kernel Density Estimate - KDE



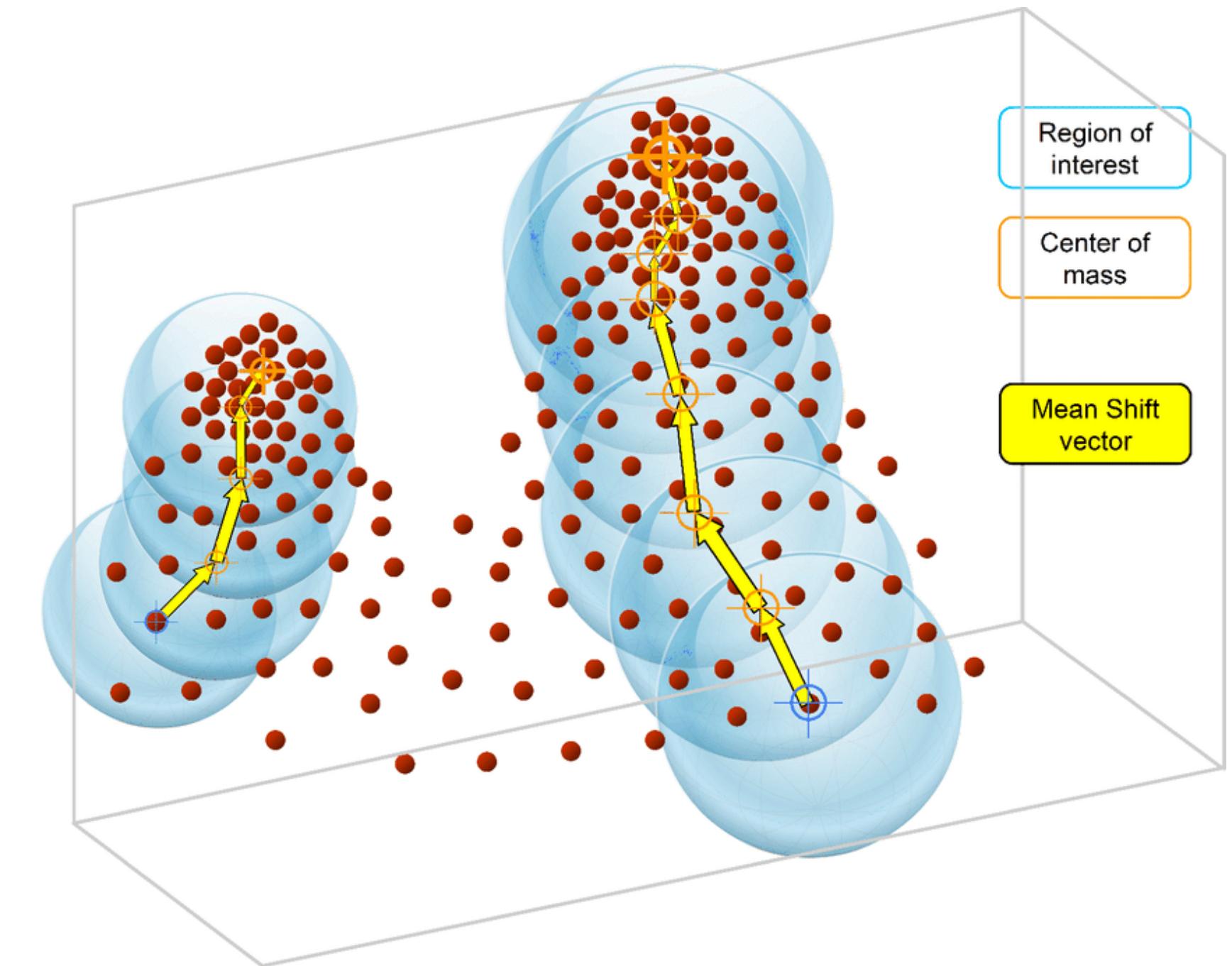
## Kernel Density Estimate - KDE

Bandwidth của kernel tác động  
lên độ “mịn” của đường KDE



*Bandwidth khác nhau cho đường  
KDE khác nhau - Nguồn: Wikipedia*

# Ước lượng Gradient và Mean Shift vector

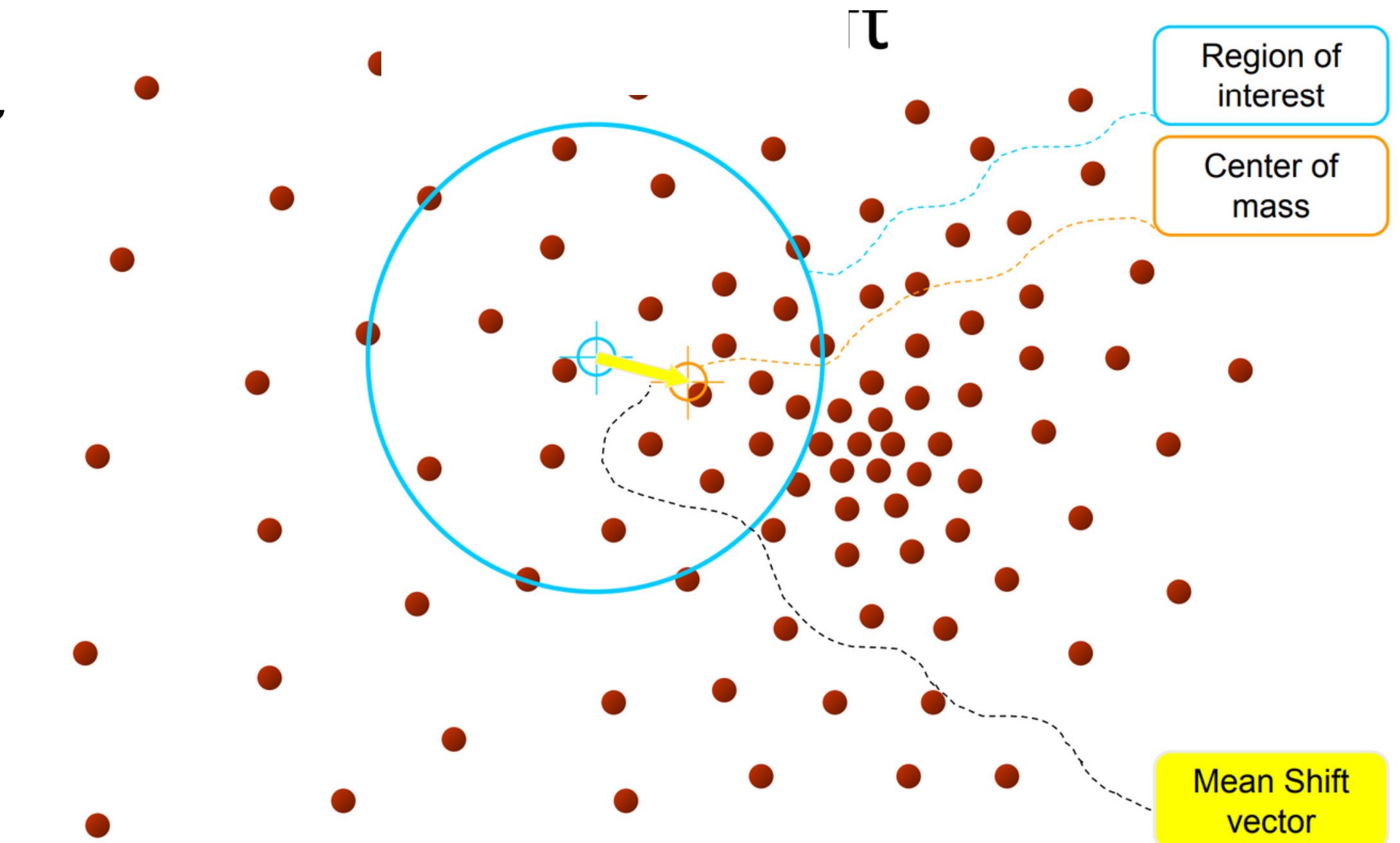


## Bước 1: Khởi tạo kernel cho từng điểm dữ

liệu:

Bắt đầu từ mỗi điểm dữ liệu, ta tạo kernel cho điểm đó. Nếu biểu diễn các điểm dữ liệu trong không gian 2 chiều, xét đường tròn với trọng tâm là điểm dữ liệu đó, và bán kính bằng với độ rộng bandwidth.

Đường tròn này gọi là window giúp kiểm soát phạm vi tìm kiếm.

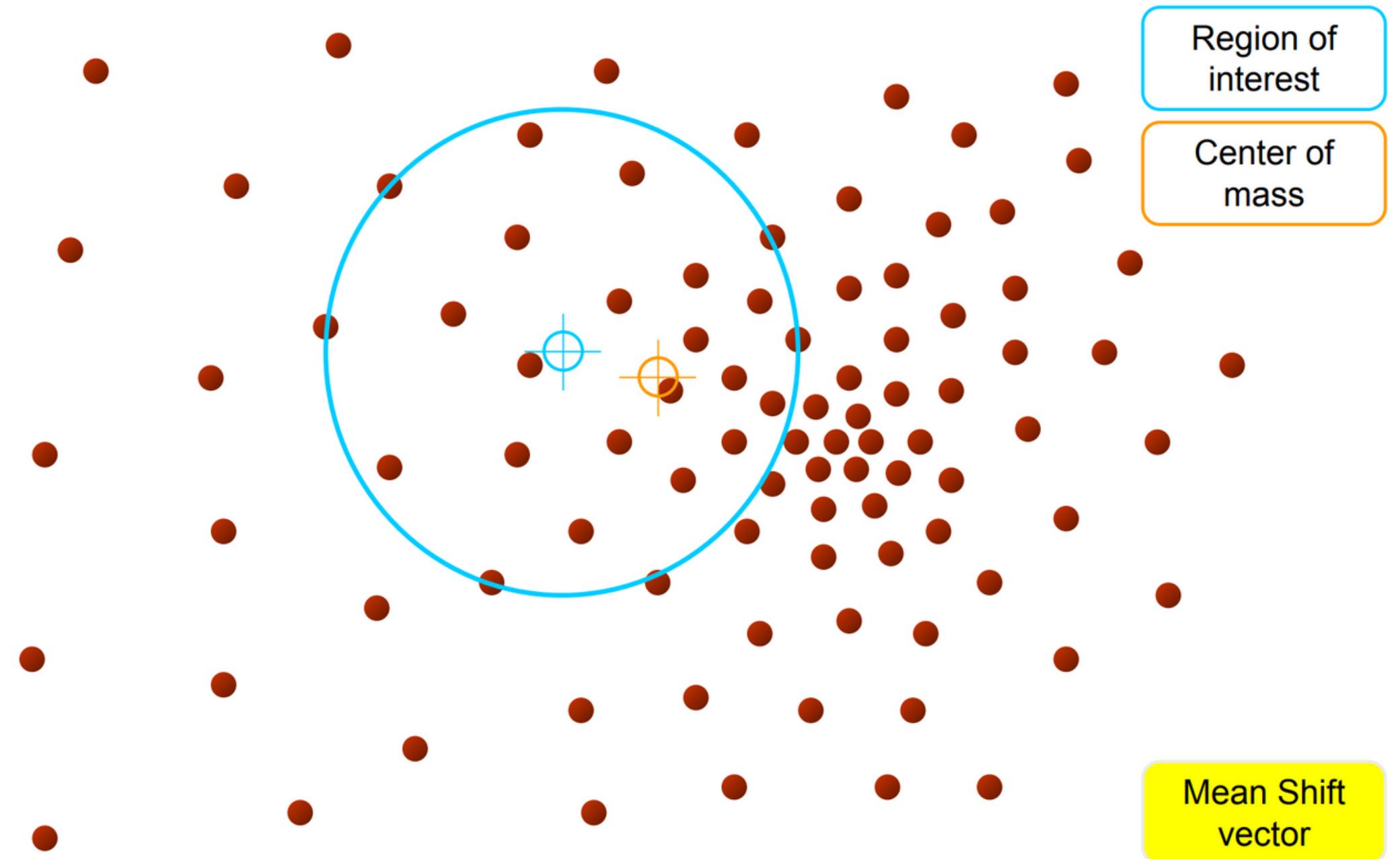


[Y. Ukrainitz & B. Sarel]

## Bước 2: Tính toán centroid:

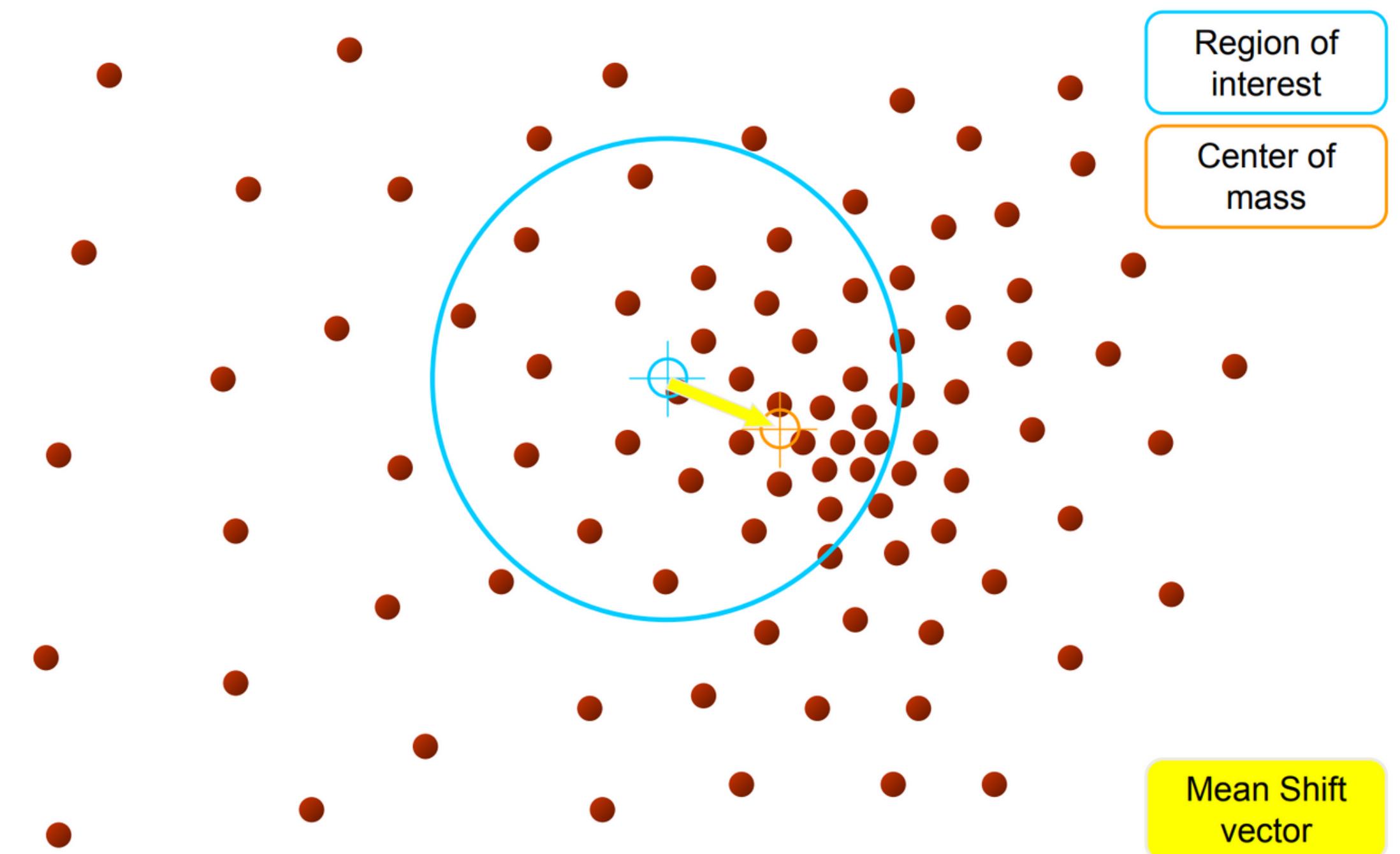
Tính trọng tâm mới từ các điểm nằm trong window.

Nếu tạo ra một trọng tâm giống với trọng tâm vừa có thì tiến tới Bước 4.



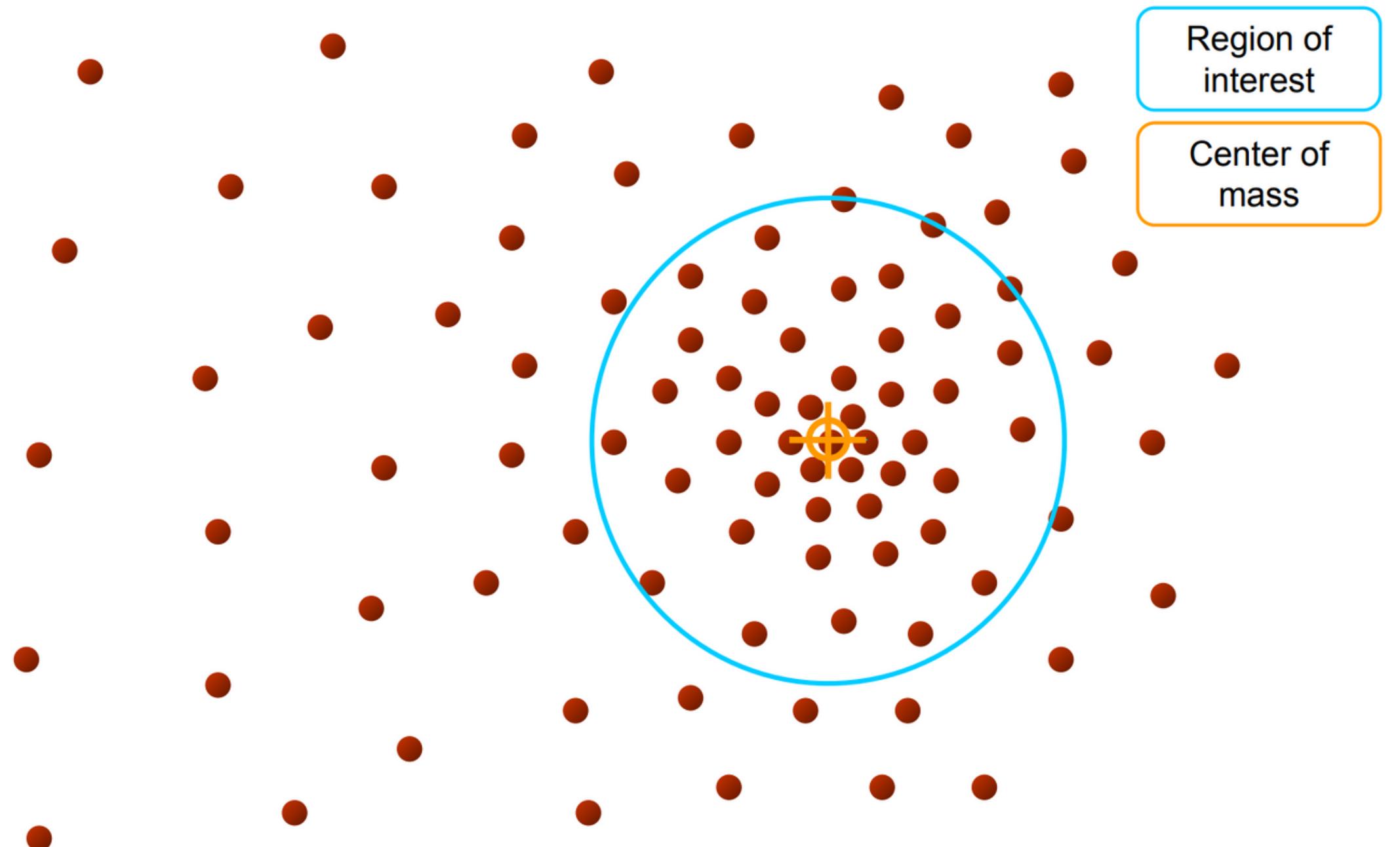
[Y. Ukrainitz & B. Sarel]

**Bước 3: Di chuyển window đến vị trí mới,** với trọng tâm mới được tính ở bước 2. Quá trình trọng tâm di chuyển đến trọng tâm mới tạo ra vector Mean Shift, vector này mô tả hướng và độ lớn giữa 2 trọng tâm, giúp window dần tiến đến các vùng dày đặc. Sau đó lặp lại bước 2.



[Y. Ukrainitz & B. Sarel]

**Bước 4:** Trả về các trung  
tâm cụm cuối cùng và  
phân hoạch dữ liệu vào  
các cụm.



[Y. Ukrainitz & B. Sarel]

## ƯU ĐIỂM

- Không cần xác định trước số cụm.
- Thuật toán không nhạy cảm đối với các điểm ngoại vi và đảm bảo hội tụ.
- Mean Shift không đặt ra giả định nào về hình dạng của các cụm dữ liệu

## NHƯỢC ĐIỂM

- Việc chọn bandwidth lại không dễ dàng.
- Không hiệu quả với không gian đặc trưng có số chiều lớn.

## Đồ Án KTHP Môn học Máy Học

# TỔNG QUAN BỘ DỮ LIỆU

8 thuộc tính

2000 quan sát

STT	Tên thuộc tính	Mô tả	Kiểu dữ liệu
1	CustomerID	Mã định danh cho mỗi khách hàng để phân biệt (2000 khách hàng)	int64
2	Gender	Giới tính của khách hàng: nam hoặc nữ	object
3	Age	Tuổi của khách hàng khi thu thập dữ liệu	int64
4	Annual Income	Thu nhập hàng năm của khách hàng (đơn vị: đô la)	int64
5	Spending Score	Điểm số dựa trên hành vi mua sắm của khách hàng	int64
6	Profession	Nghề nghiệp của khách hàng	object
7	Work Experience	Số năm làm việc của khách hàng	int64
8	Family Size	Số lượng thành viên trong gia đình của khách hàng	int64

Đồ Án KTHP Môn học Máy Học

# TIỀN XỬ LÝ DỮ LIỆU

# Xử lý dữ liệu bị thiếu

```
df.isna().sum()
```

```
Gender      0
Age         0
Annual Income 0
Spending Score 0
Profession   35
Work Experience 0
Family Size   0
Education    0
```

```
#Xóa dòng có missing data
df.dropna(inplace=True)
```

# Xử lý outliers

```
#Xem thông tin các dòng có chứa outliers
```

```
df[df['Work Experience']==17]
```

	Gender	Age	Annual Income	Spending Score	Profession	Work Experience	Family Size
392	Male	21	119116	30	Artist	17	4
405	Female	65	119889	11	Artist	17	6
473	Male	20	130813	92	Artist	17	5
566	Female	19	180331	14	Artist	17	5
603	Female	91	69720	78	Lawyer	17	6

```
#Loại ra những người bắt đầu đi làm khi dưới 16 tuổi
```

```
df.drop(df[df['Age']-df['Work Experience']<16].index, axis = 0,inplace=True)
```

Đồ Án KTHP Môn học Máy Học

**ÁP DỤNG THUẬT TOÁN MEAN SHIFT  
XÂY DỰNG MÔ HÌNH PHÂN CỤM CHO  
BỘ DỮ LIỆU SHOP CUSTOMER DATA**

# CHỈNH DẠNG DỮ LIỆU

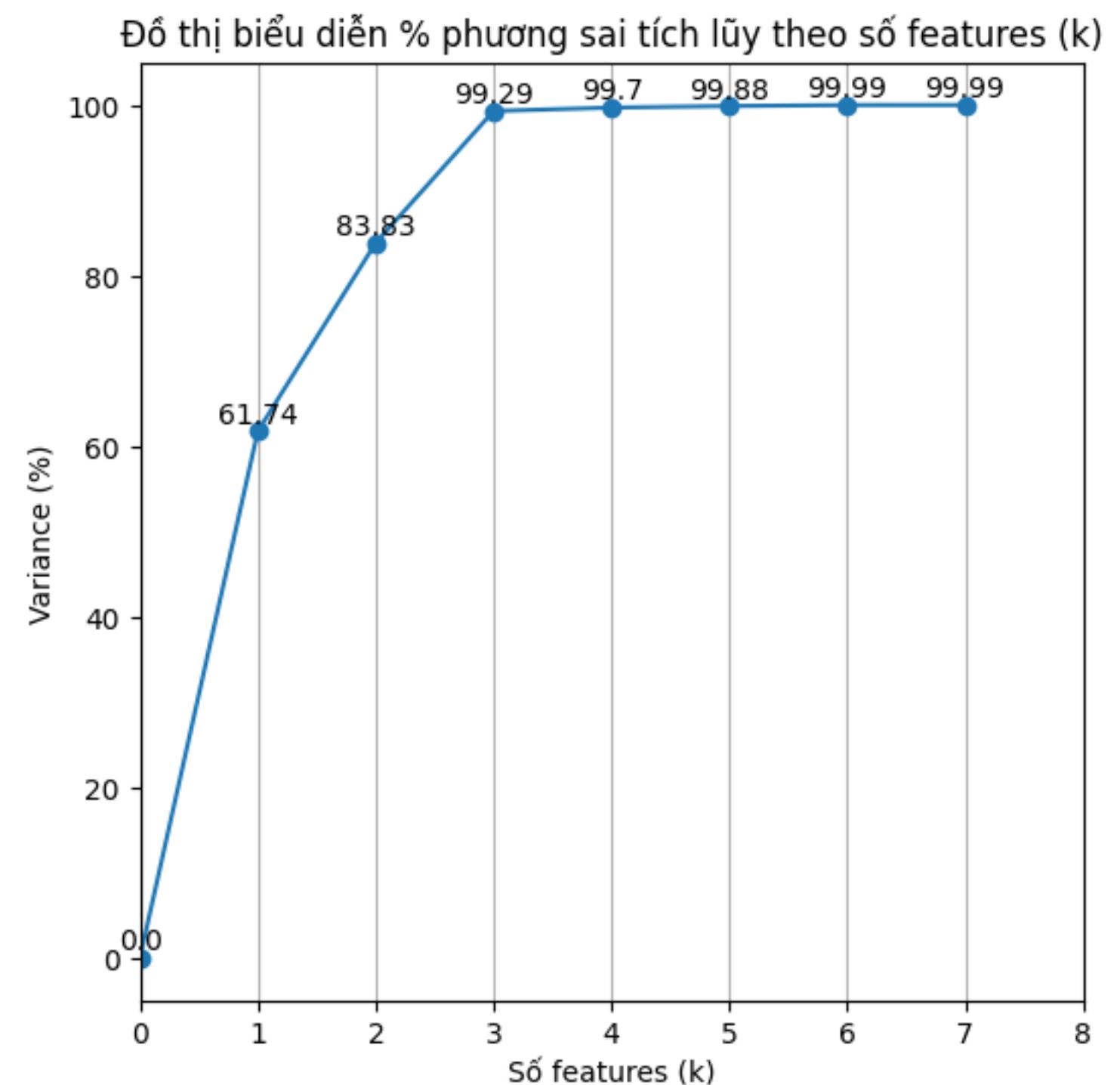
- LABEL ENCODING
- CHUYỂN ANNUAL INCOME VỀ THANG K\$

	Gender	Age	Annual Income	Spending Score	Profession	Work Experience	Family Size	
0	1	19	15.0	39	5		1	4
1	1	21	35.0	81	2		3	3
2	0	20	86.0	6	2		1	1
3	0	23	59.0	77	7		0	2
4	0	31	38.0	40	3		2	6

# GIẢM CHIỀU DỮ LIỆU PCA

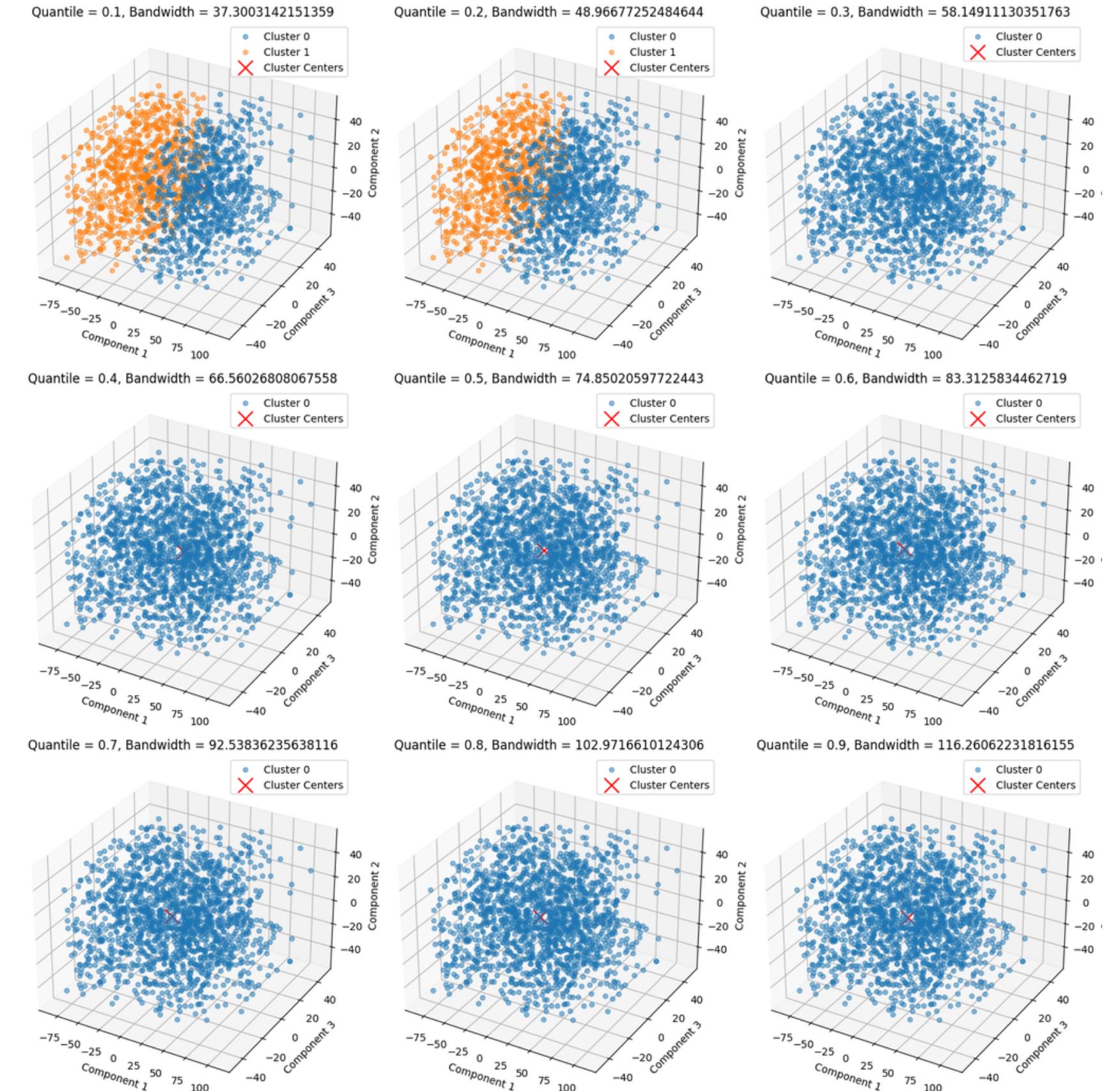
Xem biểu đồ % phương sai tích lũy theo k

=> Chọn điểm gãy k = 3

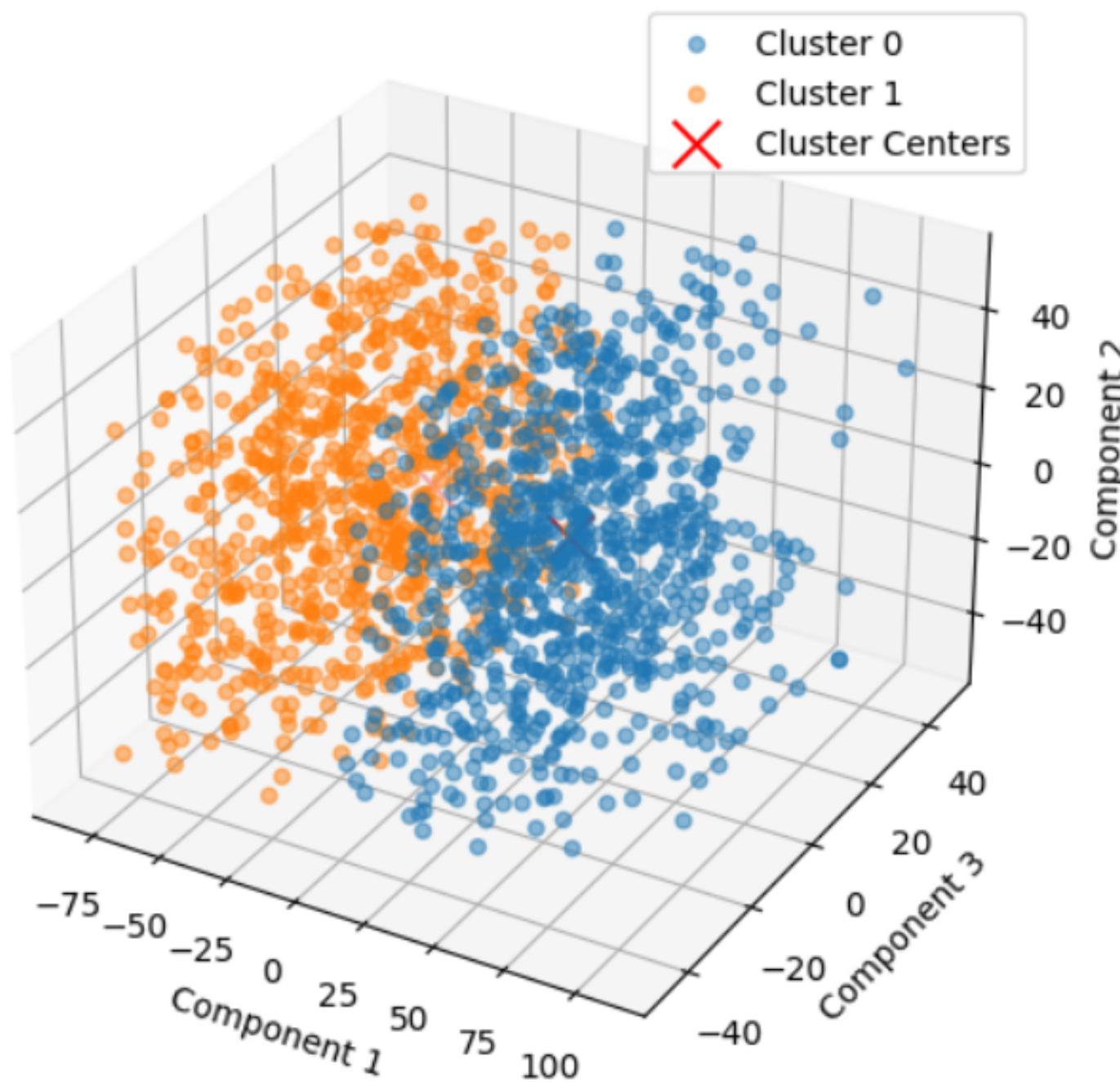


# XÂY DỰNG MÔ HÌNH MEAN SHIFT

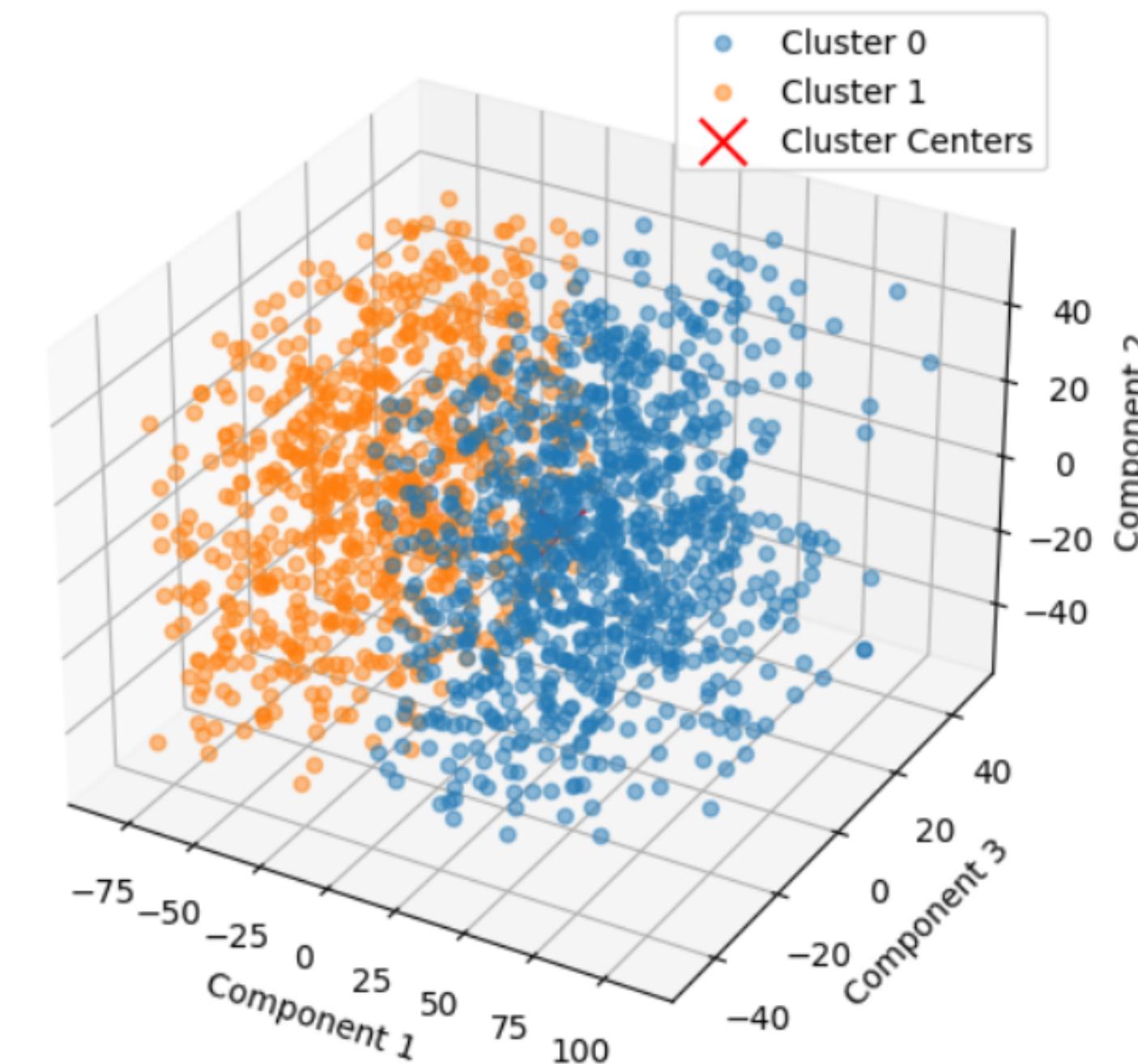
**1. Lựa chọn bandwidth**  
dựa trên kết quả biểu diễn  
gom cụm



Quantile = 0.1, Bandwidth = 37.3003142151359



Quantile = 0.2, Bandwidth = 48.96677252484644



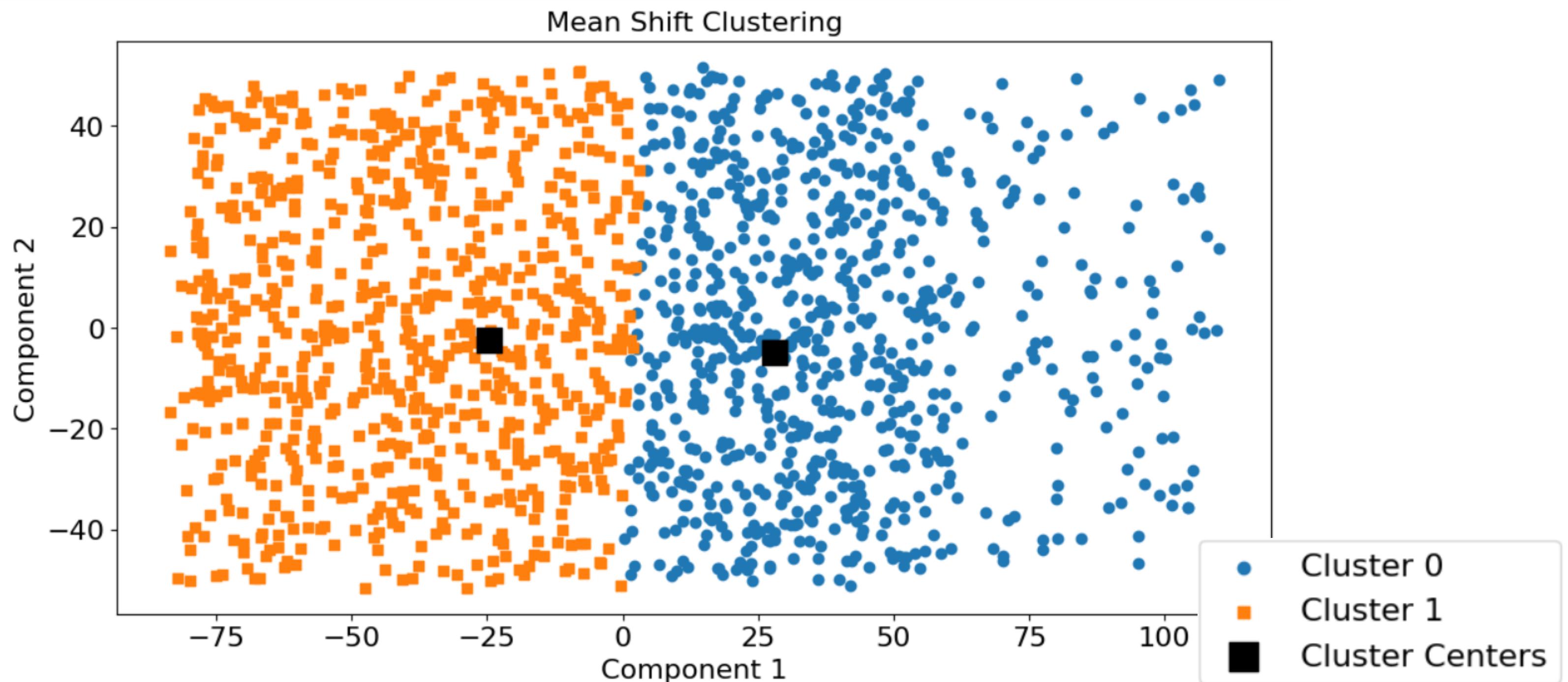
# XÂY DỰNG MÔ HÌNH MEAN SHIFT

## 2. Lựa chọn bandwidth dựa trên phương pháp Grid Search

Best Params: { 'bandwidth': 37.3003142151359}

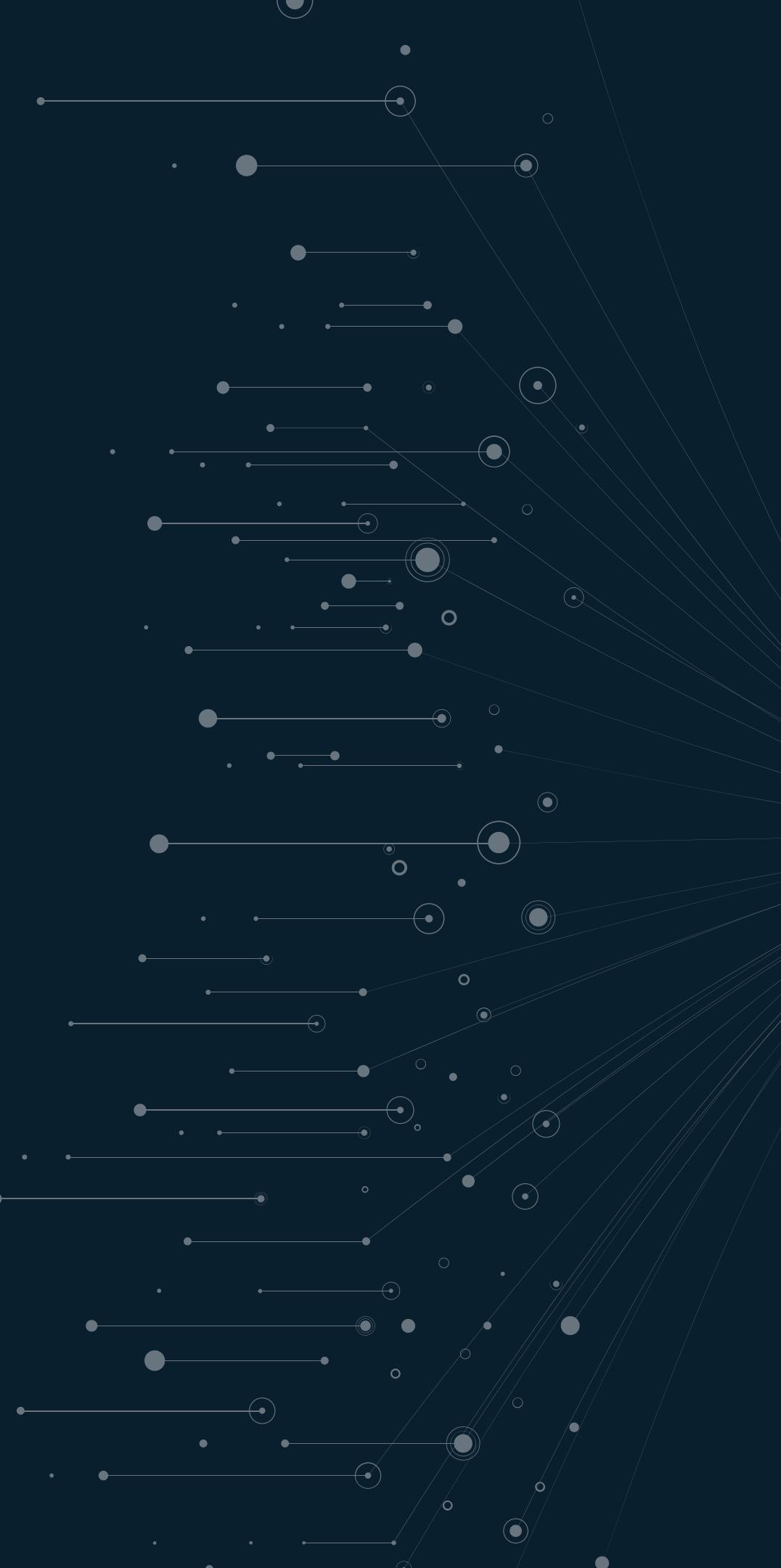
Avg Score - Silhouette Score: 0.3695521952867356

## 3. Biểu diễn kết quả phân cụm dưới dạng 2D

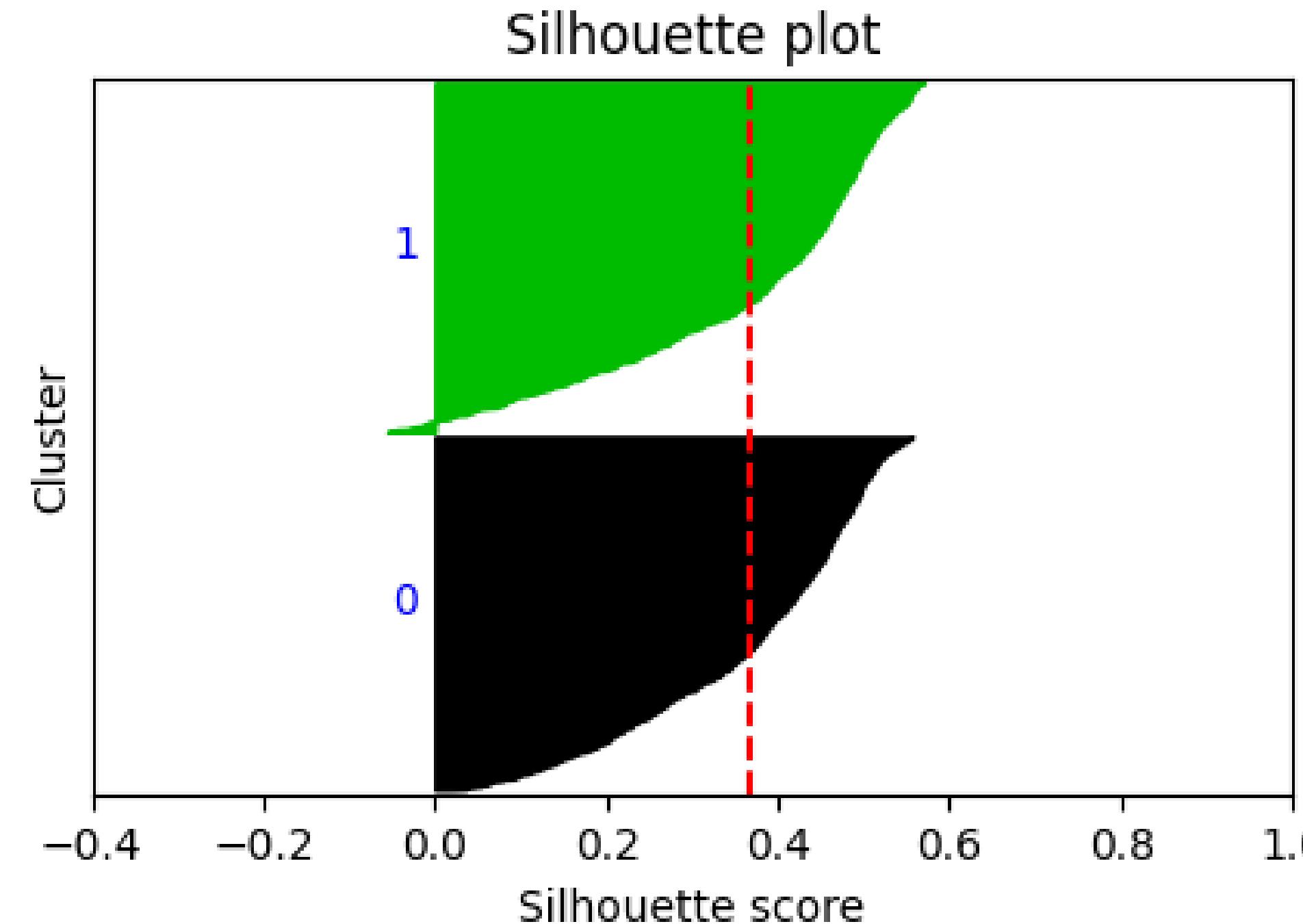


Đồ Án KTHP Môn học Máy Học

# ĐÁNH GIÁ



# ĐÁNH GIÁ PHÂN CỤM BẰNG ĐIỂM SILHOUETTE



# THANK YOU