TUNIS BUSINESS SCHOOL
UNIVERSITY OF TUNIS

# Finding Peaks in Data: A Dive into Mean Shift Clustering

Aymen MSADDAK

**Instructor:** Dr. Monstassar BEN MESSAOUD

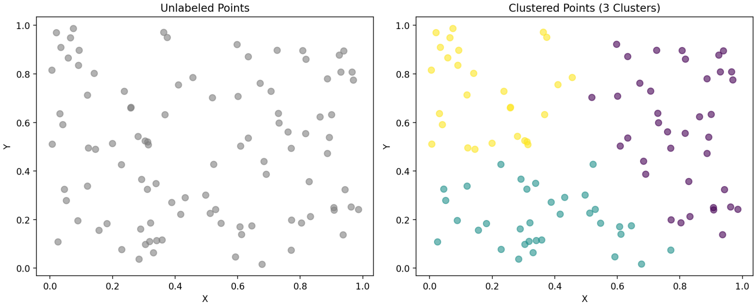
Tunis Business School

IT 430 Machine Learning

# Outline

# Introduction - What is Clustering?

- Clustering is a form of **unsupervised learning** that groups data points into **meaningful clusters** based on their **similarity**.
- Clustering = trying to discover *hidden structures in your data*.



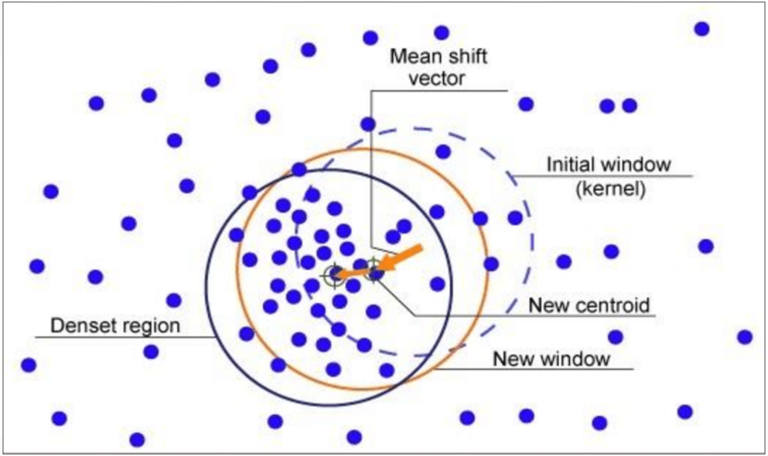
# Why do we Need Mean Shift? – K-Means Limitations

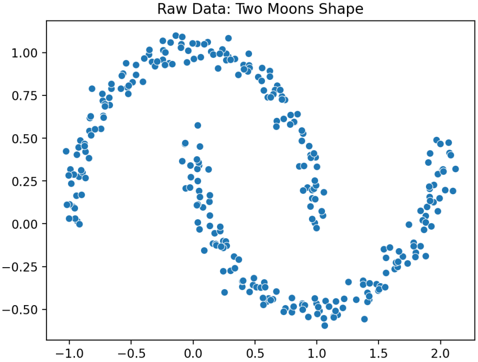
- Requires a **pre-defined number of clusters (k)**, which is not always clear.
- Assumes clusters are **spherical** and **similar in size**.
- Struggles with **non-linear** or **irregularly shaped data**.
- Sensitive to **initial placement of centroids**.

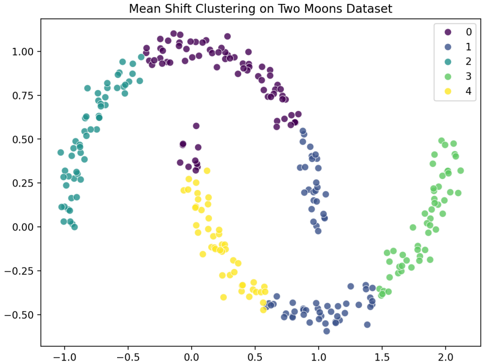
*K-Means tries to "force" the data into k round shapes — even if when data isn't round or evenly spaced.*

# Philosophy Behind Mean Shift Clustering

- A **"hill climbing" process** over the data's **probability density function (PDF)**.
- Each point "looks around" itself in a small window (a circular neighborhood).
- It computes the **mean position** of all the nearby points inside that window.
- Then it *shifts* toward that mean.
- All points **move uphill** to the nearest peak of density.
- Each peak (mode) becomes a **cluster center**.






Mean Shift follows the "moon-shaped" density curves naturally, grouping points by where they're most concentrated.

# Non-Parametric Clustering

- "Non-parametric" means it doesn't assume a **fixed number of clusters** or a particular shape.
- Clusters emerge naturally from the data's underlying distribution.
- Mean Shift ideal for **exploratory analysis**.
- The goal is to **discover structure**, not enforce it

| **Situation** | **Why Mean Shift Helps** |
|---|---|
| Unknown the number of clusters | It automatically estimates clusters |
| Clusters are non-spherical or uneven | It adapts to data density |
| Detecting natural patterns | It follows data's distribution |
| Image segmentation | Each color/intensity mode becomes a cluster |
| Exploring small/medium datasets | Easy to visualize and interpret |

# Background

## Local Density Estimation around a point $x$ [1]

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- Dataset: $X = \{x_1, x_2, \ldots, x_n\}, \quad x_i \in \mathbb{R}^d$
- $K(x)$: measures proximity
- $h$: the bandwidth controlling the neighborhood radius

# Background

## Mean Shift Vector [2]

$$m(x) = \frac{\sum_{i=1}^{n} x_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)} - x$$

The Mean Shift vector is the direction and distance the point should move to reach the local mean

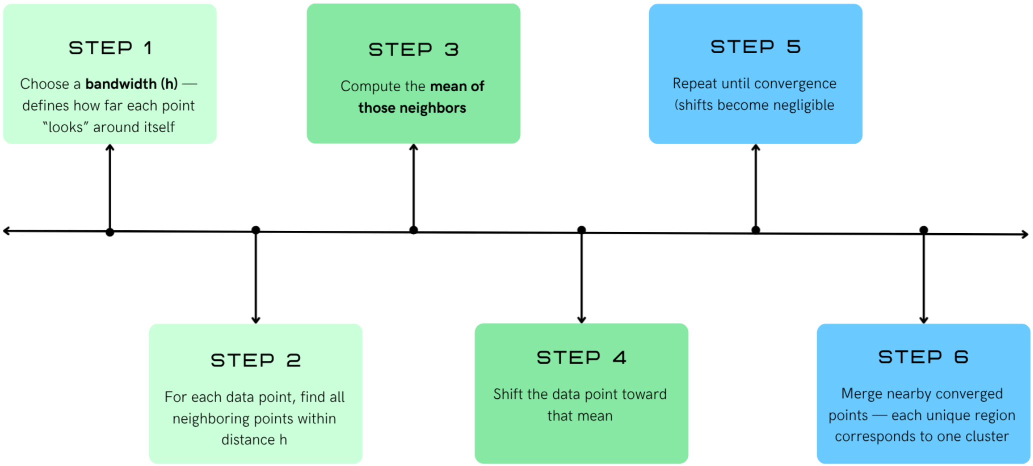
Interpretations:

- The numerator: the weighted mean of neighboring points.
- The subtraction $(-x)$: gives the *shift direction* toward higher density.
- If $m(x) = 0$, the point is at a local maximum (mode).

So each iteration moves $x$ to: $x_{\text{new}} = x_{\text{old}} + m(x_{\text{old}})$

And we repeat this process until convergence.

# Mean Shift Algorithm Overview



# The Role of the Bandwidth (h)

- The **most critical hyperparameter** in Mean Shift [3].
- **The kernel size** determines the neighborhood radius, it is used to calculate local density and shift points toward higher-density regions.
- $K(x) = e^{-\frac{\|x\|^2}{2h^2}}$

| **Bandwidth Size** | **How Mean Shift Reacts** |
|---|---|
| Small $h$ | Captures fine details — possibly **too many clusters** |
| Large $h$ | Smooths over fine details — possibly **too few clusters** |
| Just right | Distinguishes meaningful clusters clearly |

# Metrics for Evaluating Clusters

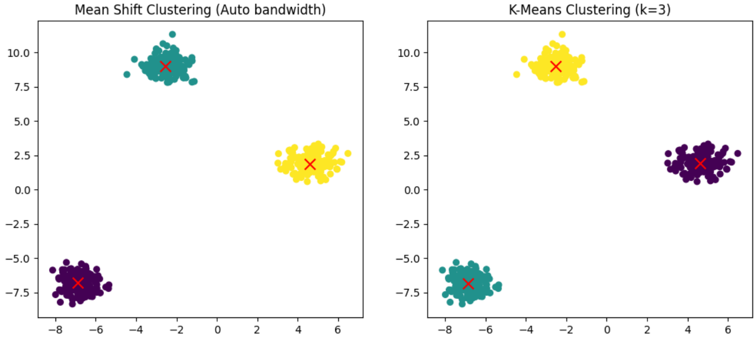
| Metric | What It Measures | Interpretation |
|---|---|---|
| **Silhouette Score** | Average distance between points in the same cluster vs. points in different clusters | Closer to 1 $\rightarrow$ better defined clusters |
| **Davies–Bouldin Index** | Ratio of within-cluster distance to between-cluster separation | Lower is better |
| **Calinski–Harabasz Index** | Ratio of between-cluster dispersion to within-cluster dispersion | Higher is better |

# Research Questions

- "*RQ1*: How does Mean Shift identify clusters without knowing their number in advance?"
- "*RQ2*: How sensitive is Mean Shift to the bandwidth parameter?"
- "*RQ3*: When does Mean Shift outperform K-Means?"
- "*RQ4*: What are the limitations of Mean Shift for large or high-dimensional datasets?"
- "*RQ5*: How does Mean Shift handle noise and outliers?"

# RQ1 — How does Mean Shift identify clusters without knowing their number in advance?

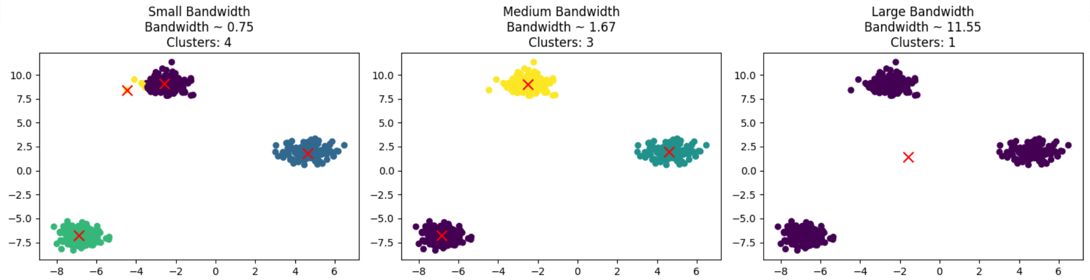
- A simple dataset with three blob-shaped clusters.
- Mean Shift followed **data density peaks**.
- Mean Shift **automatically** detected **3 clusters**.
- With K-Means, k had to be **explicitly set to 3**.
- Mean Shift is powerful for exploratory scenarios where the structure is unknown in advance.



# RQ2 : How sensitive is Mean Shift to the bandwidth parameter?

- Bandwidth $h$ determines the **radius of density influence**:
  - Acts like a **resolution knob**
  - Small $h \rightarrow$ fine-grained detection (risk of over-clustering)
  - Large $h \rightarrow$ broad smoothing (risk of merging clusters)
- Bandwidth is set via *estimate_bandwidth()* function.
- Bandwidth is the **most critical hyperparameter** in Mean Shift.
- Controls the trade-off between over-clustering and over-smoothing.
- *The goal is to find a balanced bandwidth that captures actual structure without noise.*

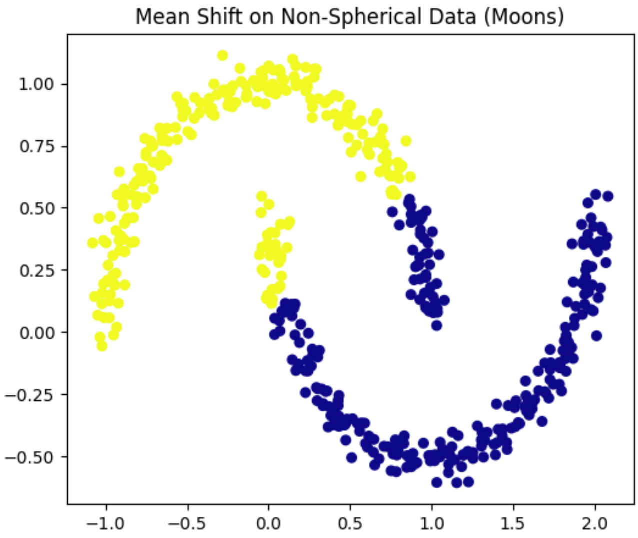
# RQ2 : How sensitive is Mean Shift to the bandwidth parameter?




As discussed previously, bandwidth strongly affects Mean Shift's cluster detection resolution.

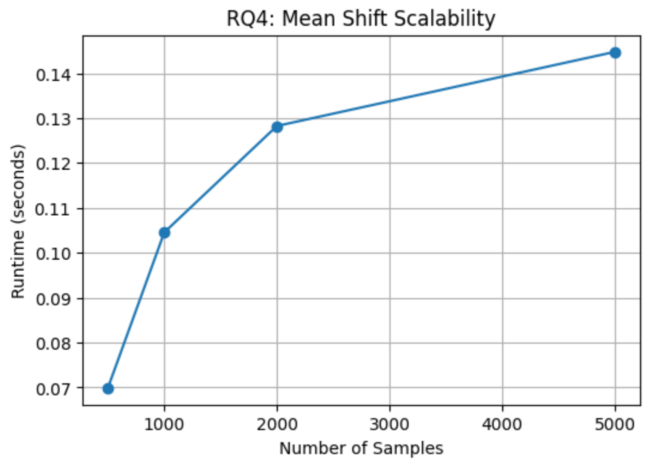
# RQ3 : When does Mean Shift outperform K-Means?

- Moons Dataset
- Mean Shift accurately identified the **two curved clusters**.
- K-Means incorrectly divided the data into **convex segments**.
- Mean Shift is highly effective for **irregular**, **curved**, or **complex cluster structures**.



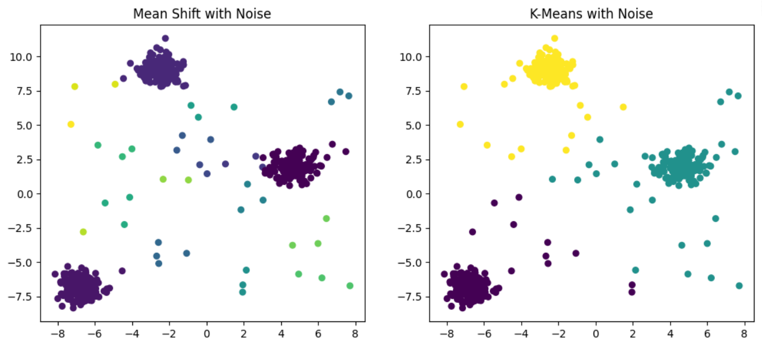
# RQ4 : What are the limitations of Mean Shift for large or high-dimensional datasets?

- Increasing sample sizes showed that **runtime increases sharply** as dataset size grows.
- Cluster accuracy stayed **consistent** (3 clusters correctly detected).
- Mean Shift performs distance computations for all points, leading to **quadratic complexity** $O(n^2)$
- Mean Shift offers **high clustering accuracy** but its main limitation is **scalability**.



# RQ5 : How does Mean Shift handle noise and outliers?

- Mean Shift **ignored scattered noise points** (low-density regions).
- K-Means **forced noise points into clusters**, reducing cluster quality.
- Mean Shift is **resilient to noise**, which enhances clustering reliability.



# Detecting Crime Hotspots in Sfax Using Mean Shift Clustering

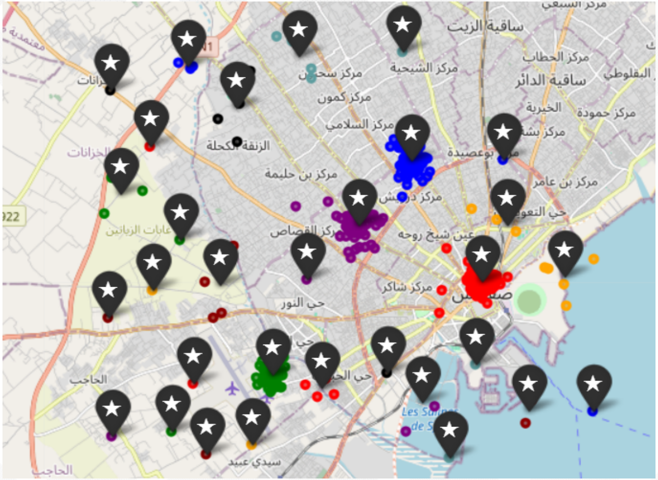
- Identify **crime concentration zones** in **Sfax**, Tunisia for urban security analytics.
- Granular incident-level crime data for Tunisia is not publicly available.
- Simulate incident coordinates around **real neighborhoods** known for **mixed commercial, residential, and transit activity.** [4]
- Areas modeled:
  - Sfax Medina
  - Sakiet Ezzit
  - Thyna
  - El Ain
- Random noise incidents were added to mimic scattered, low-density events.
- **Features**: Latitude, Longitude
- **Bandwidth Selection**: estimate_bandwidth()
- Python implementation uses scikit-learn's MeanShift API [5].
- **Radius is set to 3** for each crime point so the map remains readable.

# Detecting Crime Hotspots in Sfax Using Mean Shift Clustering

## Key Findings

- Mean Shift **automatically detected** multiple crime hotspots.
- Clusters form only in **truly dense areas**.
- Scattered crime events did not create false clusters.
- Hotspots appear around **city center**, **transport hubs**, and **residential pockets**.
- Helps authorities focus patrol units in priority zones.
- Reduces over-policing in low-crime areas.

# Detecting Crime Hotspots in Sfax Using Mean Shift Clustering




*Click to open interactive crime hotspot map (Sfax)*

# Conclusion

Mean Shift Clustering is a powerful **non-parametric technique** that discovers clusters based on **data density**, eliminating the need to pre-specify the number of clusters. It excels at detecting **complex**, **non-linear cluster shapes** and shows **strong robustness to noise**, making it particularly effective for exploratory analysis and real-world applications. Its main limitation is the **computational cost for large datasets**.

Applied to Sfax crime data, Mean Shift successfully identified meaningful hotspot zones and filtered scattered incidents — demonstrating its value for geospatial and public-safety analytics.

Overall, Mean Shift is a **versatile clustering method** for **complex data**, best used when cluster structure is unknown and **precision matters more than speed**.

# References

[1] Fukunaga, K., & Hostetler, L. (1975). *Estimation of the gradient of a density function.*
Back to slide

[2] Comaniciu, D., & Meer, P. (2002). *Mean shift: A robust approach toward feature space analysis.* IEEE Transactions on Pattern Analysis and Machine Intelligence. Back to slide

[3] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Back to slide

[4] ChatGPT ⟨https://chatgpt.com⟩ Back to slide

[5] Pedregosa et al. (2011). *Scikit-learn Documentation.* Back to slide

# Thank you for your attention!

# Annex

**Probability Density Function (PDF):**
A function that describes the relative likelihood of data points in a continuous space. Mean Shift estimates the PDF to find high-density regions (cluster centers).

**Density Gradient:**
The direction and rate of the steepest increase in data density. Mean Shift follows the density gradient to move points toward the nearest high-density region (cluster center).

| **Aspect** | **K-Means** | **Mean Shift** |
|---|---|---|
| Clusters Required | Yes | No |
| Cluster Shape | Mostly spherical | Arbitrary |
| Outlier Sensitivity | Higher | Lower |
| Complexity | $O(n^2)$ | $O(n.k.d)$ |