

# CLUSTERING

# CLUSTERING



What is clustering?



Benefit of clustering



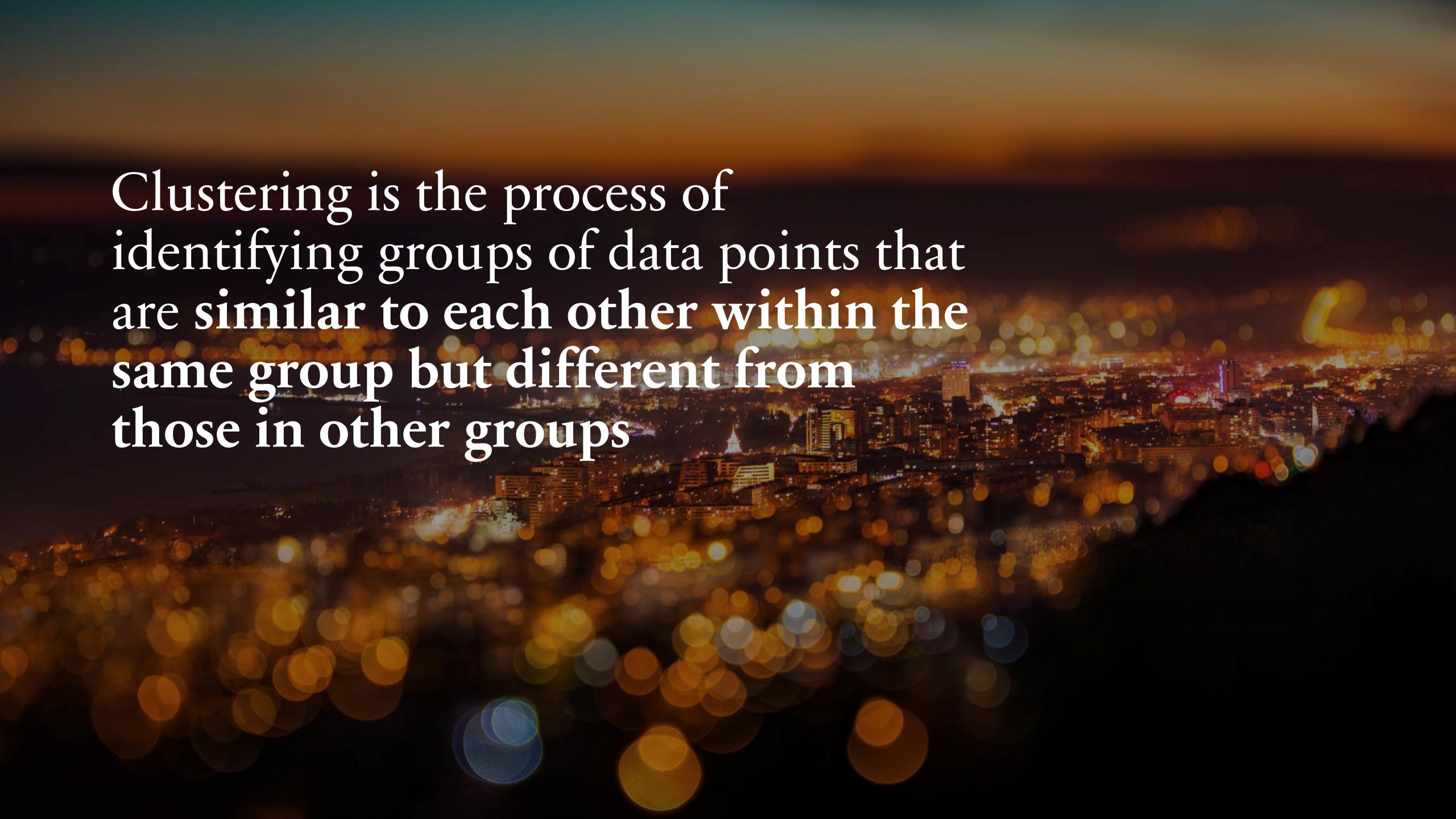
Types of clustering



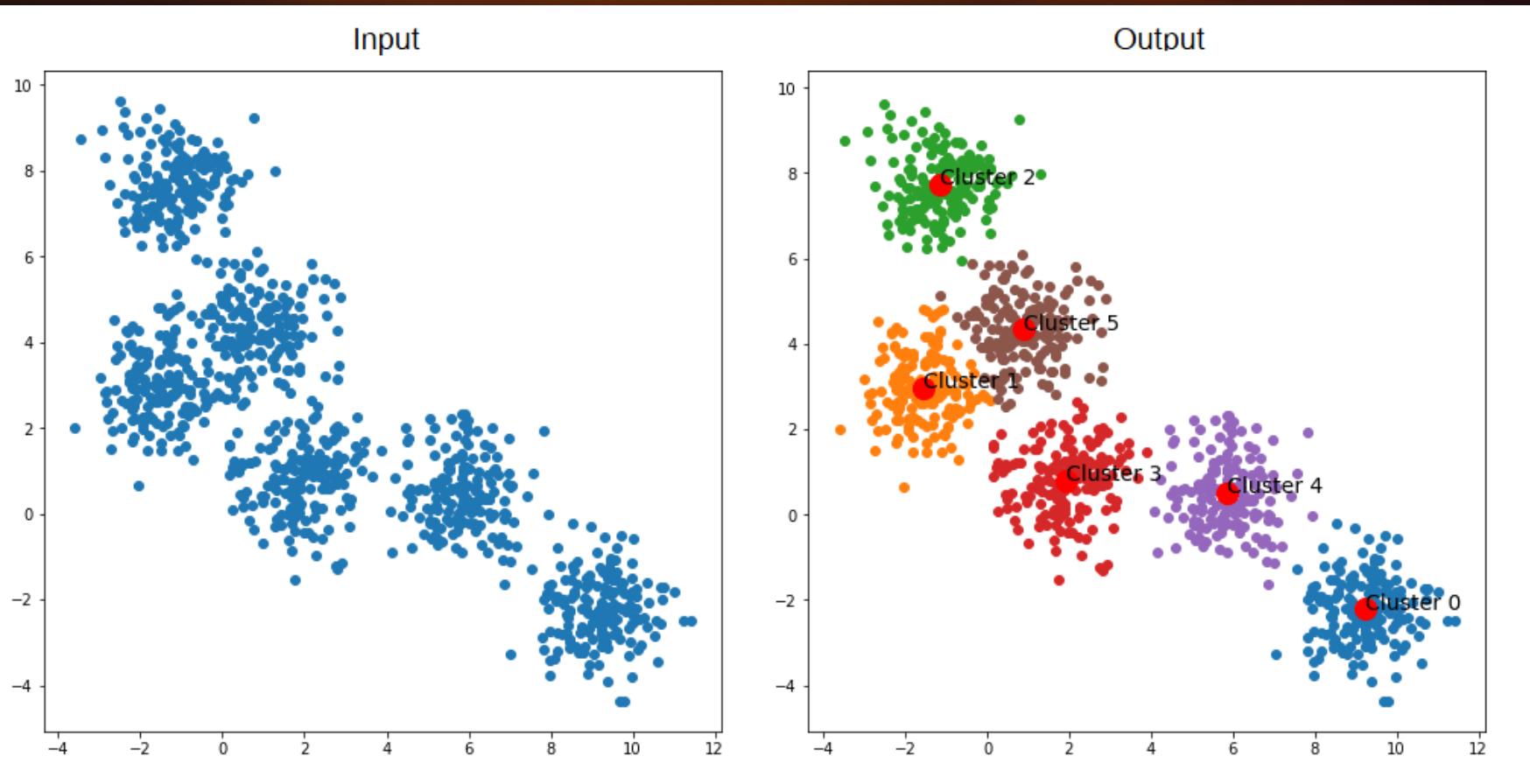
Further reading

The background of the image is a blurred night cityscape, likely a coastal city, with numerous warm-toned lights from buildings and street lamps creating a bokeh effect. The colors range from deep blues and blacks at the top to bright yellows and oranges near the horizon. A faint silhouette of a bridge or large structure is visible on the right side.

# WHAT IS CLUSTERING?



Clustering is the process of identifying groups of data points that are **similar to each other within the same group** but different from those in other groups



# BENEFIT OF CLUSTERING



Understand hidden  
patterns and structures

# Understand hidden patterns and structures

- Feature engineering
- Improve performance
- Prevent overfitting

# FOR MORE EXPLAIN

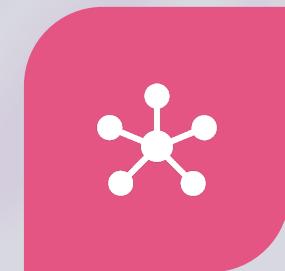
# TYPE OF CLUSTERING



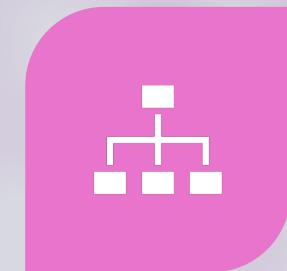
CENTROID-BASED  
CLUSTERING



DENSITY-BASED  
CLUSTERING



DISTRIBUTION-  
BASED CLUSTERING



HIERARCHICAL  
CLUSTERING

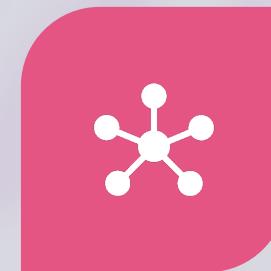
# CLUSTERING ALGORITHMS



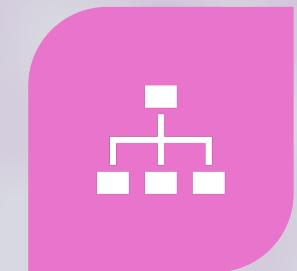
K-MEANS CLUSTERING



DBSCAN



GAUSSIAN MIXTURE  
MODEL



AGGLOMERATIVE  
HIERARCHY  
CLUSTERING

# K-MEANS CLUSTERING

A photograph of a large group of skydivers in mid-air, forming a circular pattern. The skydivers are wearing various colored jumpsuits (black, white, red, blue, yellow, green) and are arranged in concentric circles, resembling a flower or a target. The background is a clear blue sky.

# K-MEANS CLUSTERING

What is K-mean clustering?



Calculation step



Calculation example



Code



# What is K-mean clustering?

K-mean clustering is centroid-based clustering method that aims to partition  $n$  observations into  $K$  clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid)

# Calculation step

1

Choose the number of clusters K

2

Select K random datapoints from the data as centroids

3

Assign all the datapoints to the closest cluster centroid

4

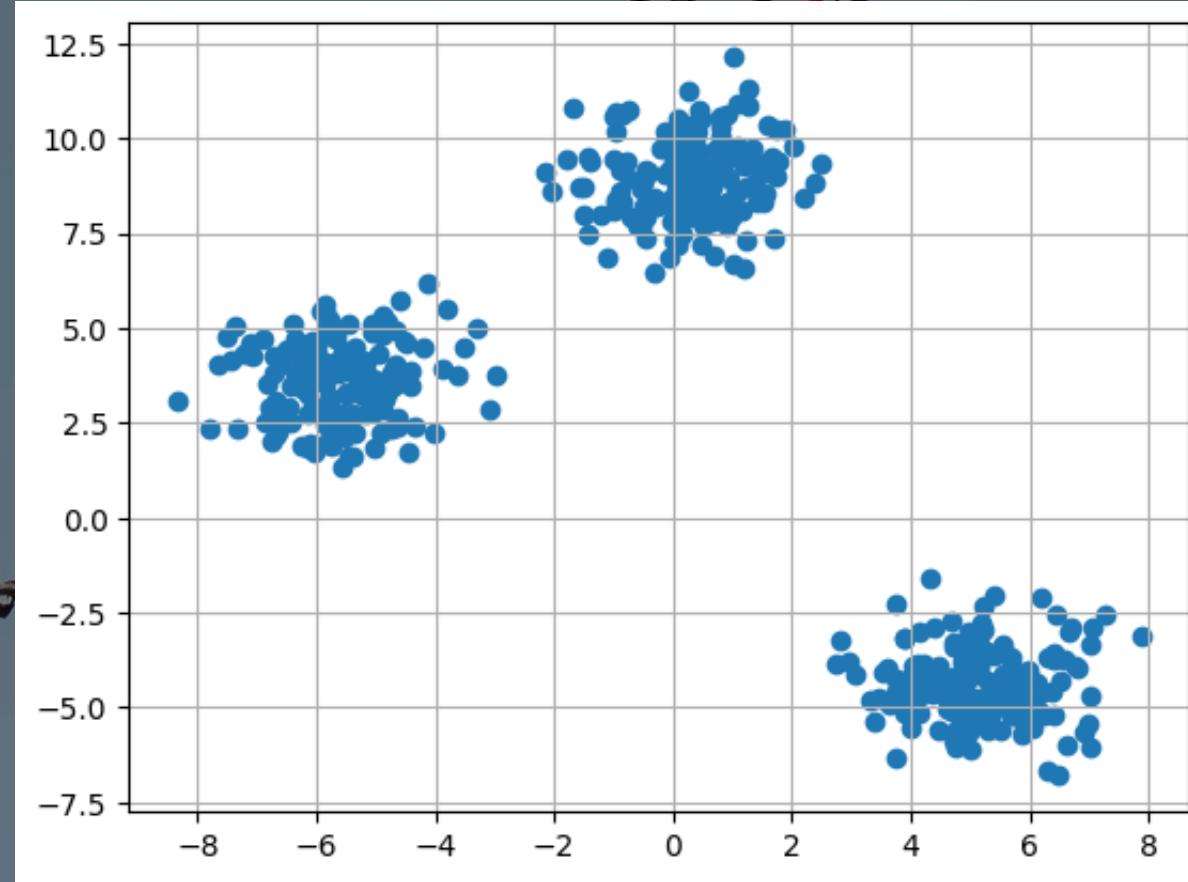
Recompute the centroids of newly formed clusters

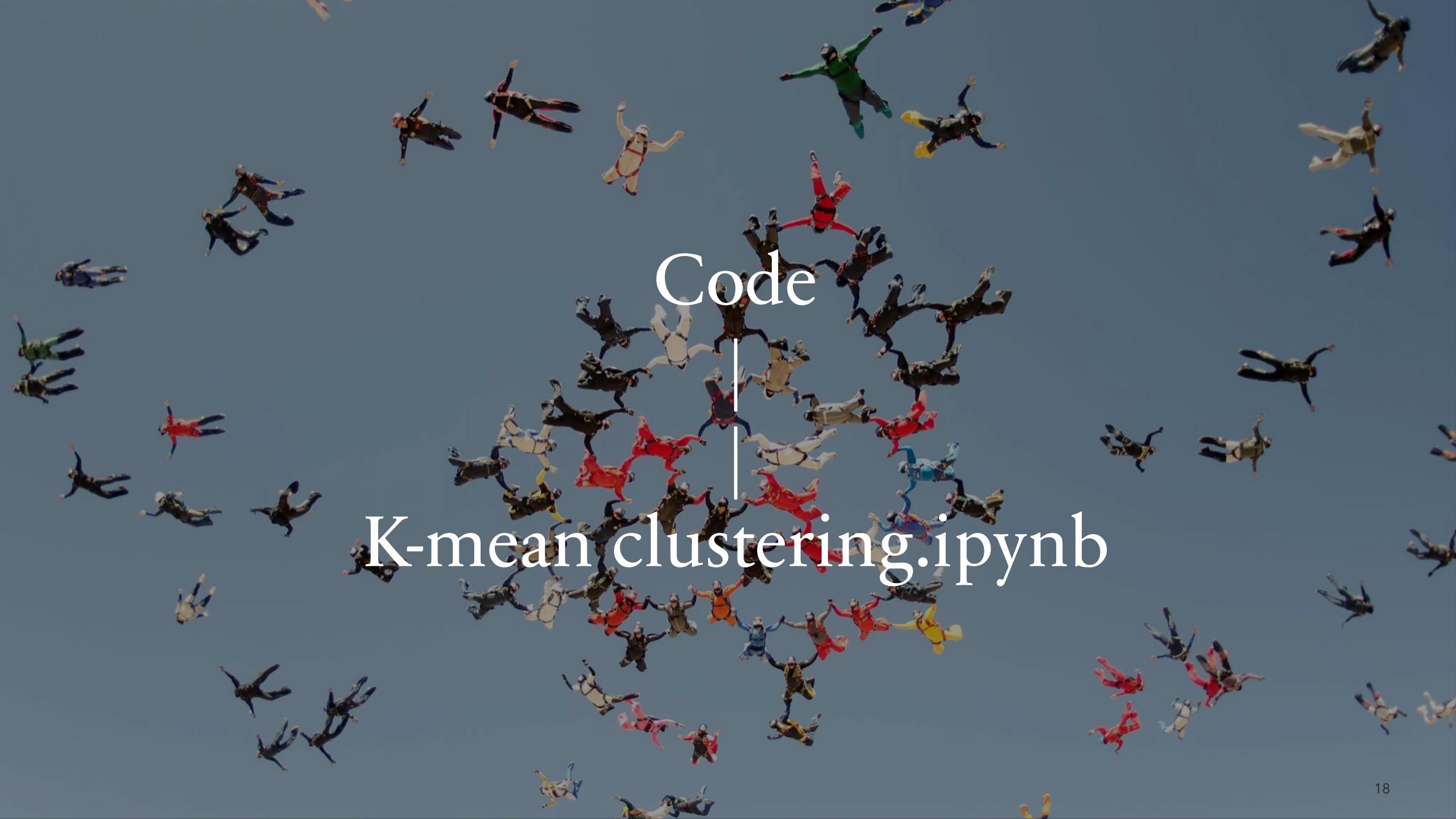
5

Repeat steps 3 and 4 until meet stopping criteria

# Calculation example

# Calculation concept





Code  
K-mean clustering.ipynb

# DBSCAN



# DBSCAN



What is DBSCAN?



Calculation Step



Calculation Example



Code

# What is DBSCAN?

DBSCAN is a density-based clustering method: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away)

# Calculation step

1

If the number of neighborhood points around  $x$  are less than  $\text{MinPts}$  and it has a core point in its neighborhood, treat it as a border point.

2

If no neighborhood points around  $x$ , treat it as a outlier.

3

Include all the density connected points as a single cluster.

4

Repeat the above steps for every unvisited point in the data set and find out all core, border and outlier points.

# Calculation step

5

Choose a value  
for eps and MinPts

6

For a particular datapoint  
(x) calculate its distance  
from every other  
datapoints.

7

For a particular datapoint  
(x) calculate its distance  
from every other  
datapoints.

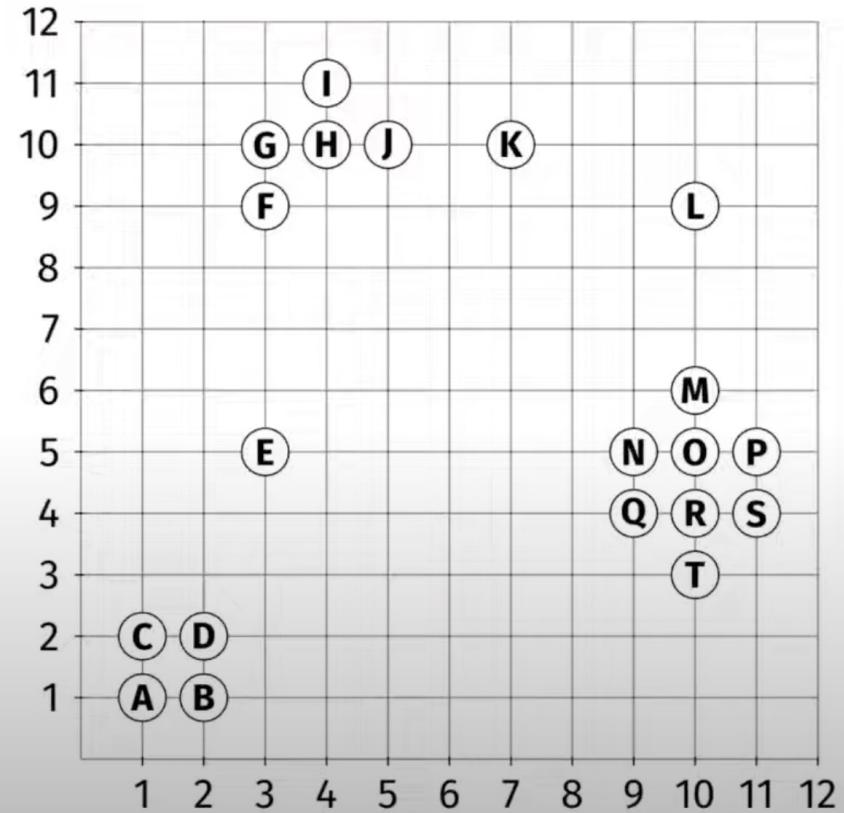
8

If the number of  
neighborhood points  
around x are greater or  
equal to MinPts then  
treat x as a core point (if  
it is not assigned to any  
cluster.)

# Calculation example

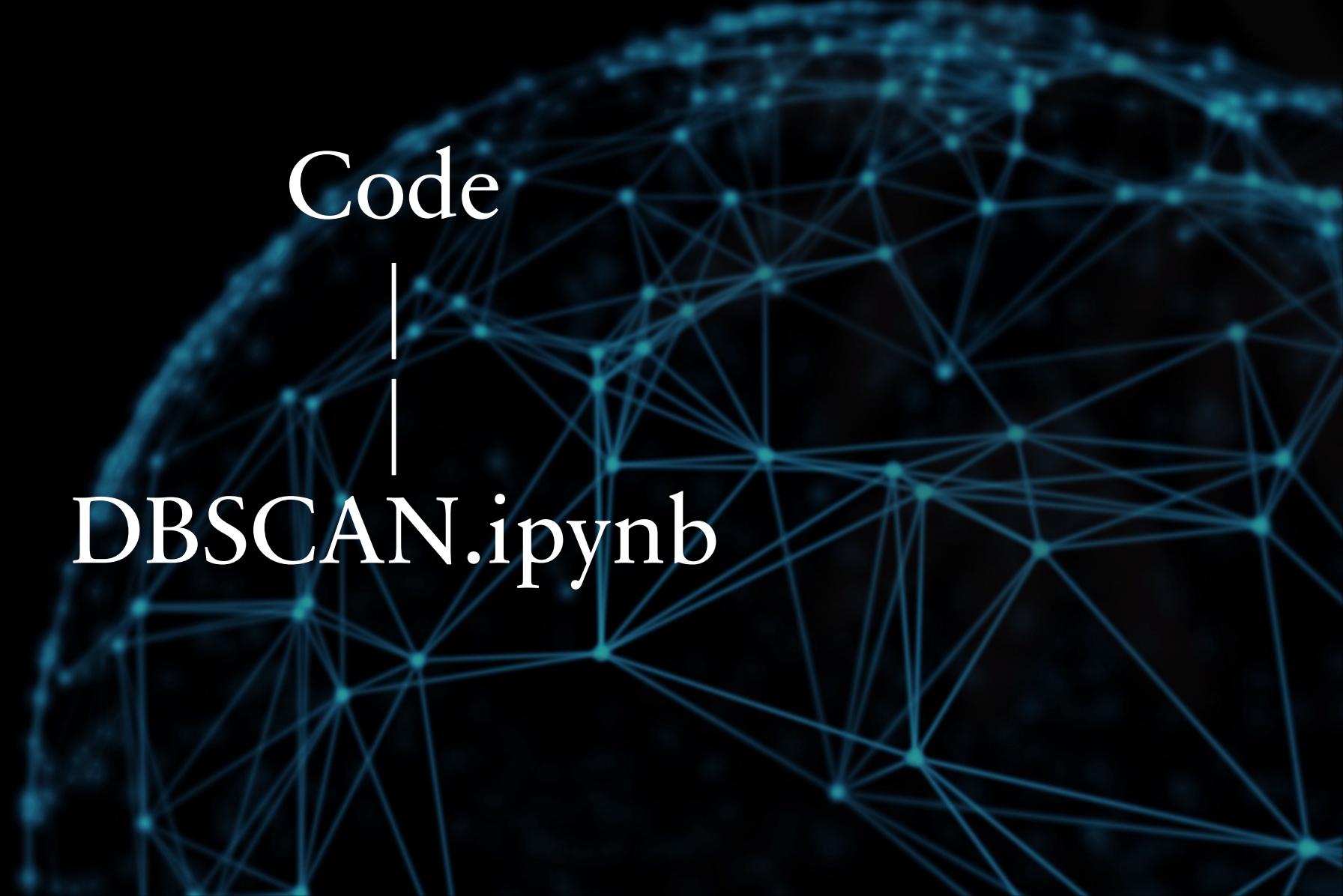


# Calculation concept



$\varepsilon = 1.75$   
minPts = 4





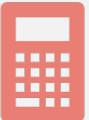
Code  
|  
DBSCAN.ipynb

# GAUSSIAN MIXTURE MODEL

# GAUSSIAN MIXTURE MODEL



What is Gaussian mixture model?



Calculation Step



Calculation Concept



Code

# What is Gaussian mixture model?

A Gaussian Mixture Model is a probabilistic model used for clustering and density estimation. It assumes that the data points are generated from a mixture of several Gaussian distributions. Each component in the mixture represents a cluster in the data, and each Gaussian distribution represents the probability distribution of data points belonging to that cluster.

# Calculation step

1

Choose the number of clusters K

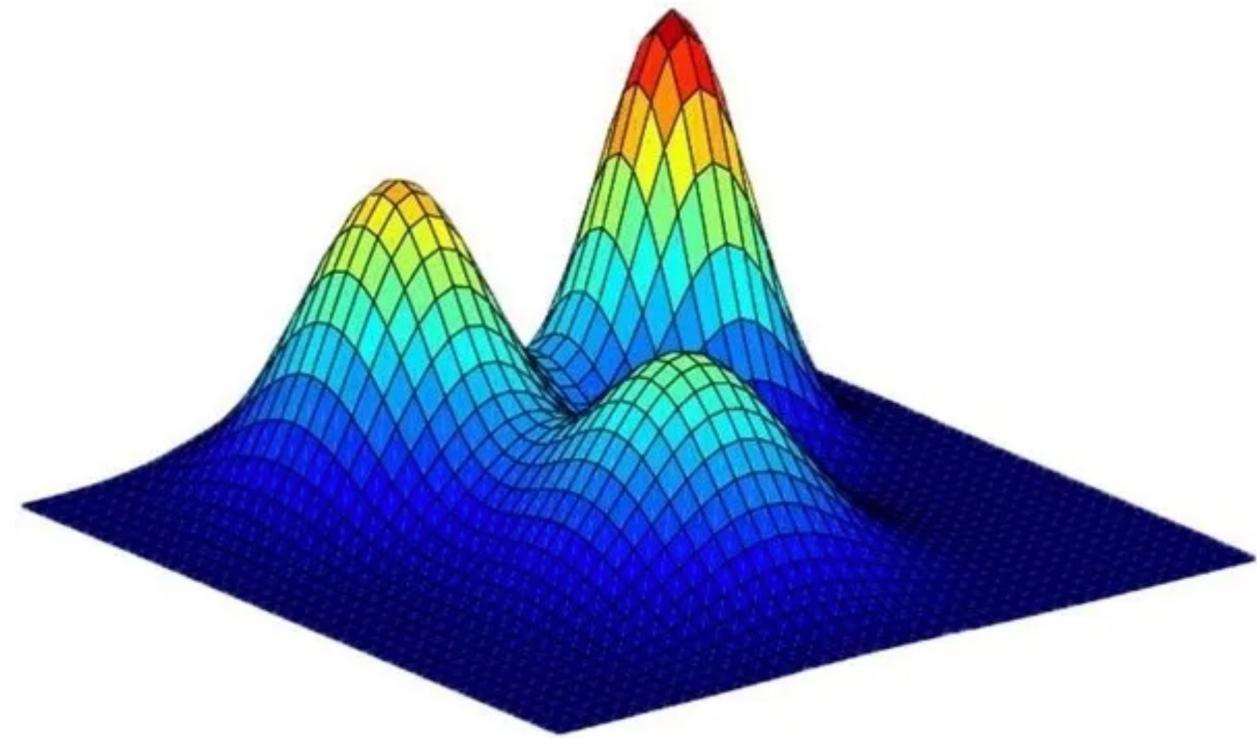
2

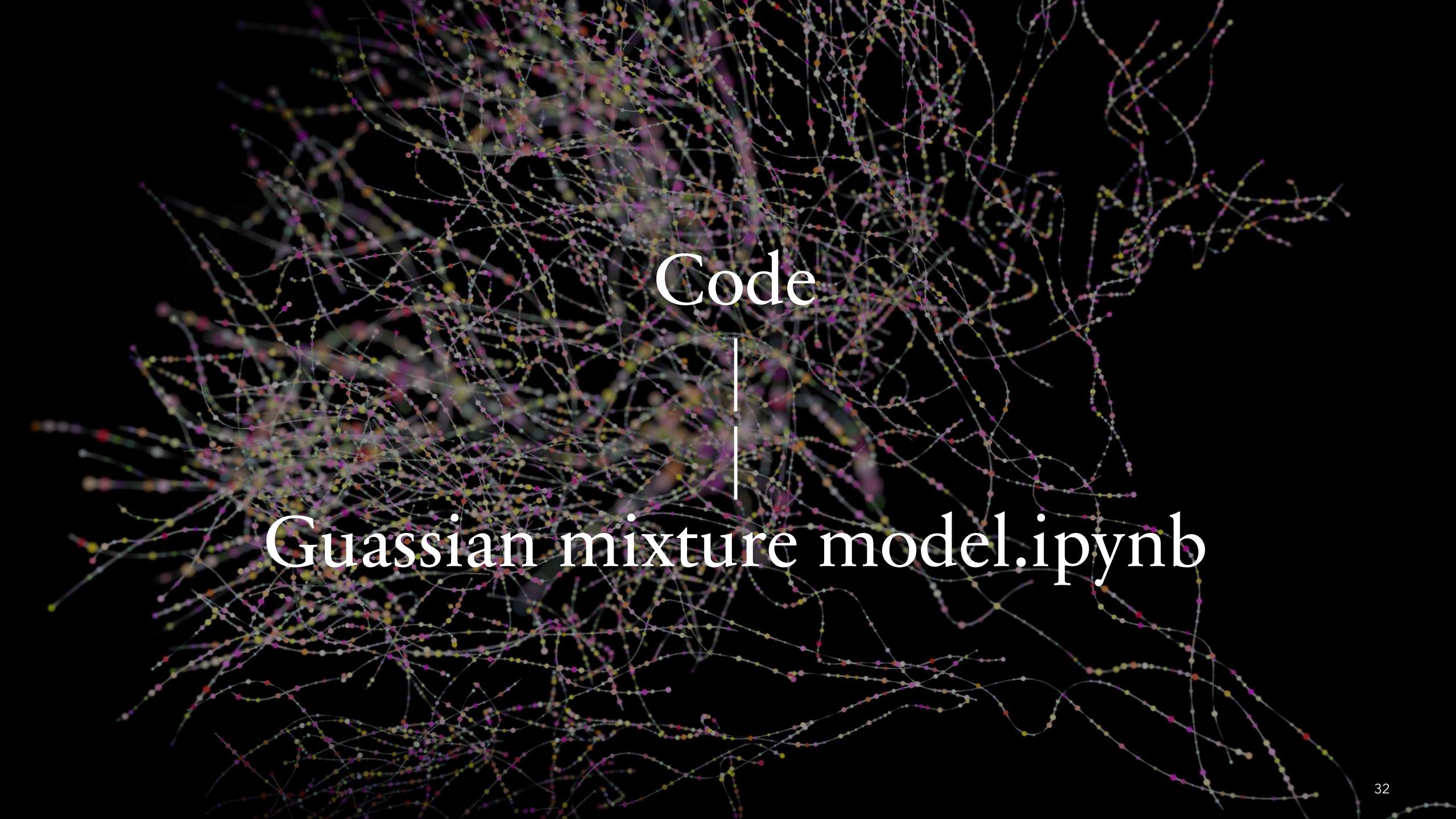
Initialize parameters  
i) means  
ii) covariances  
iii) weights

3

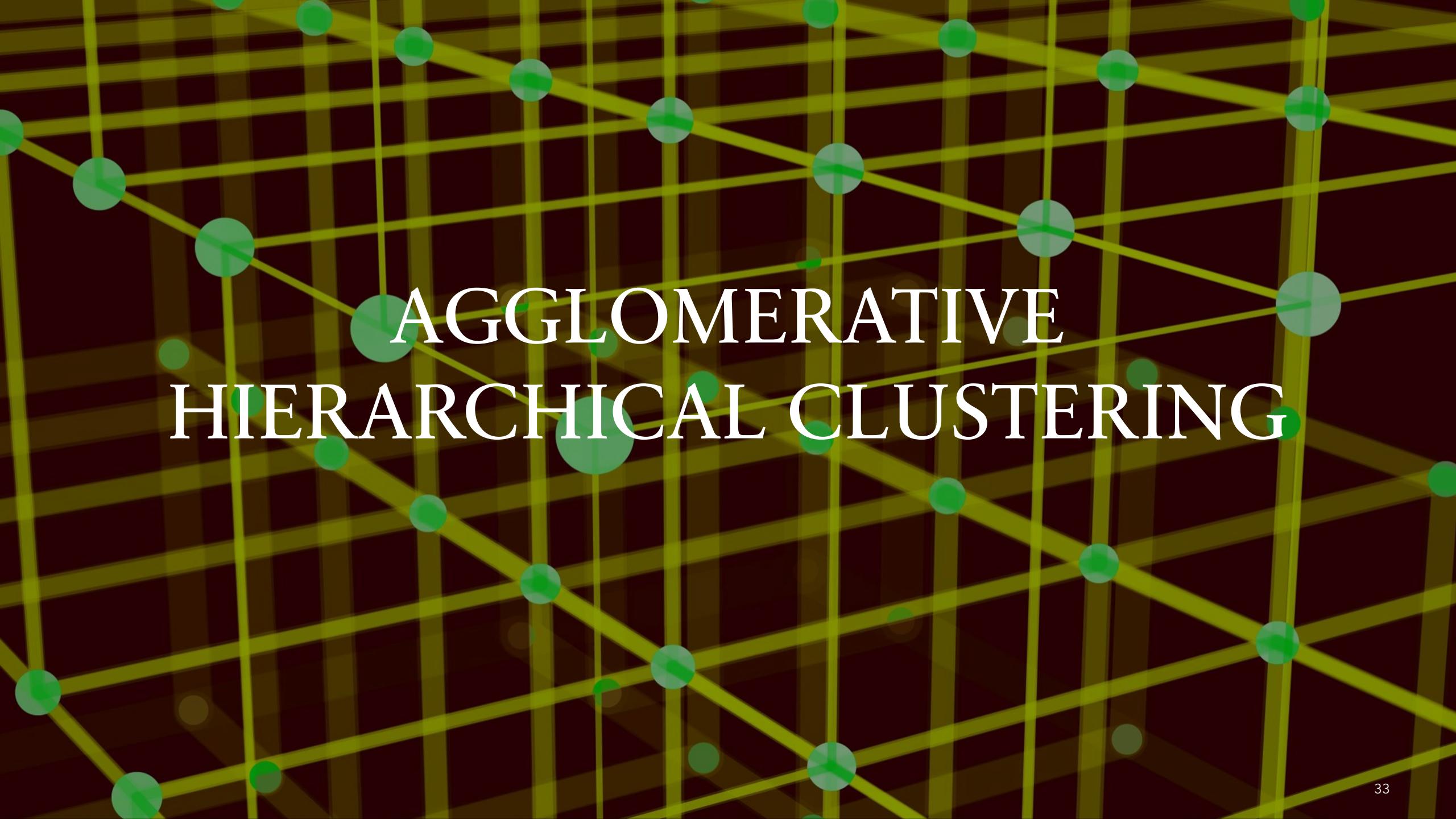
Apply EM algorithm for finding the converged parameters

# Calculation concept





Code  
|  
Guassian mixture model.ipynb



# AGGLOMERATIVE HIERARCHICAL CLUSTERING

# AGGLOMERATIVE HIERARCHICAL CLUSTERING



What is Agglomerative hierarchical clustering?



Calculation Step



Calculation Example



Code

# What is AHC?

Agglomerative hierarchical clustering (AHC) is hierarchical clustering method that seeks to build a hierarchy of clusters.

# Calculation step

1

Create each datapoint as a single cluster.

2

Take two closest datapoints or clusters and merge them to form one cluster.

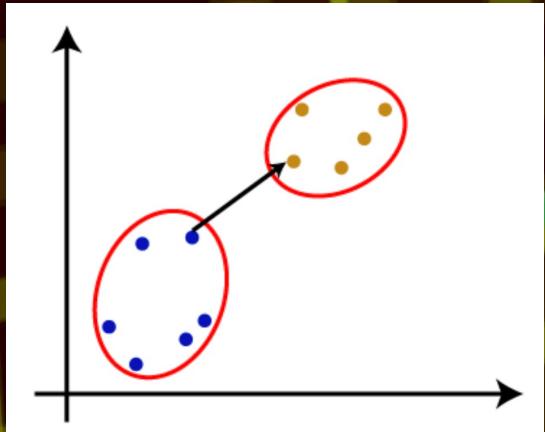
3

Repeat Step 2 until only one cluster left.

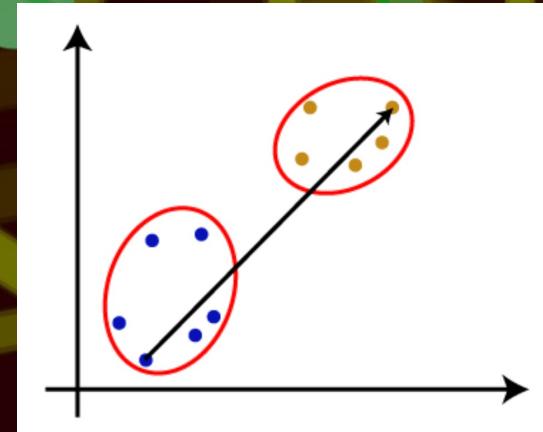
4

Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters

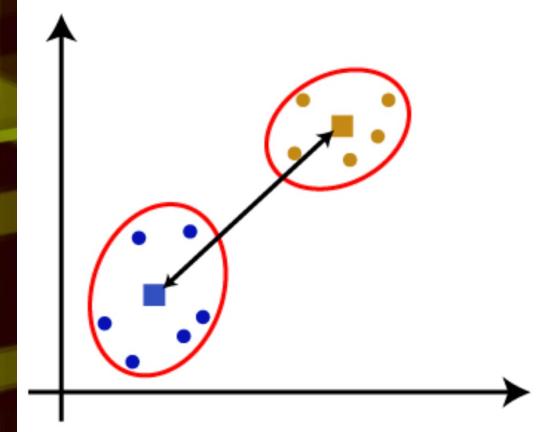
# Linkage methods



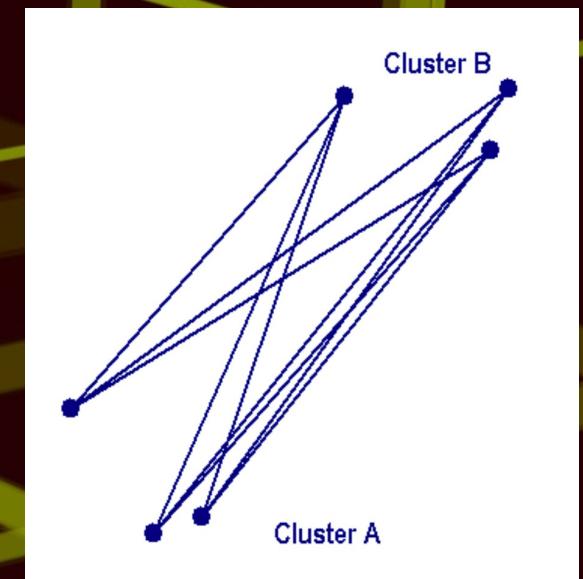
Single linkage



Complete linkage



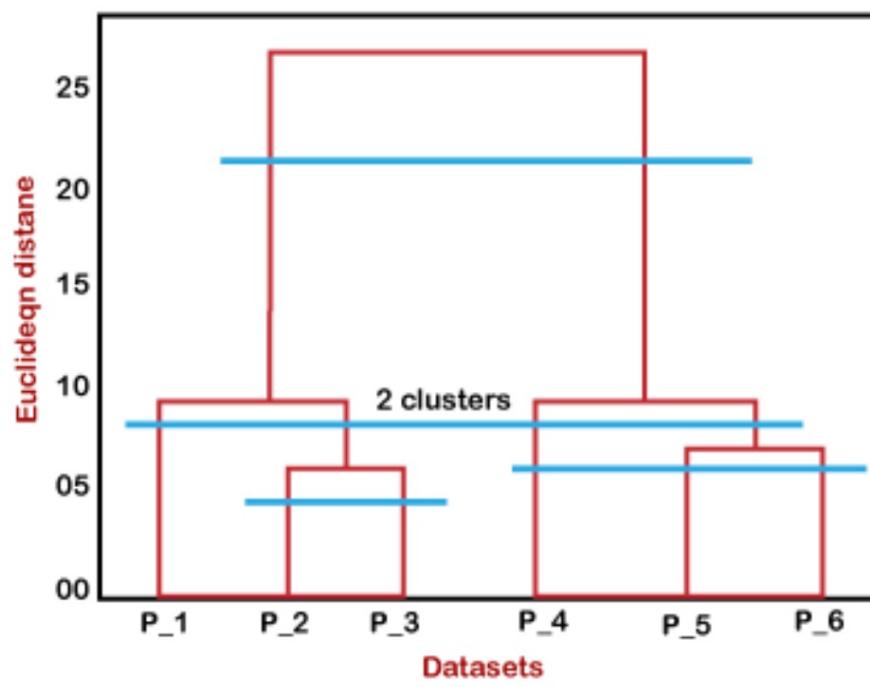
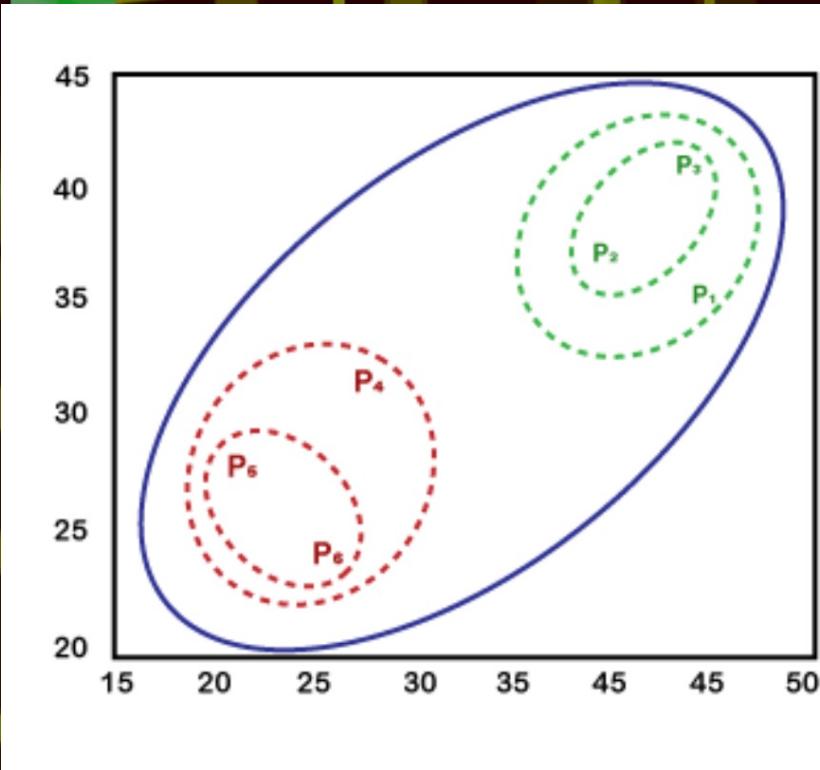
Centroid linkage



Average linkage

# Calculation example

# Calculation concept



Agglomerative hierarchical clustering.ipynb

Code

# Comparison



	<b>K-mean clustering</b>	<b>DBSCAN</b>	<b>Gaussian Mixture Model</b>	<b>Agglomerative hierarchical clustering</b>
Clustering type	Centroid-based	Density-based	Distribution-based	Hierarchical
What users define	<ul style="list-style-type: none"> <li>- Number of clusters</li> <li>- Initial cluster centroid (optional)</li> </ul>	<ul style="list-style-type: none"> <li>- Epsilon</li> <li>- MinPts</li> </ul>	<ul style="list-style-type: none"> <li>- Number of clusters</li> <li>- Initial parameters (optinoal)</li> </ul>	<ul style="list-style-type: none"> <li>- Distance threshold or number of clusters</li> <li>- Linkage method</li> </ul>
Ability to handle outliers	Sensitive to outliers	Robust to outliers	Assign low probability to outliers	Incorporate outliers into singleton clusters or ignore them
Soft clustering?	No	No	Yes	No
Suitable data shape	Sphere (convex)	Arbitary shapes (convex & non convex)	Arbitary shapes (convex & non convex) *** not that good	Arbitary shapes (convex & non convex)

# FURTHER READING

- Ensemble clustering
- Adaptive clustering
- Dynamic time wraping
- Spectral clustering
- BIRCH
- Affinity propagation clustering
- Mean-Shift clustering
- OPTICS
- Evaluation metric for clustering

# QUESTION & ANSWER



# Reference (1)

<https://pub.towardsai.net/centroid-neural-network-an-efficient-and-stable-clustering-algorithm-b2fa8cbb2a27>

[https://en.wikipedia.org/wiki/K-means\\_clustering](https://en.wikipedia.org/wiki/K-means_clustering)

<https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/>

<https://www.geeksforgeeks.org/k-means-clustering-introduction/>

<https://en.wikipedia.org/wiki/DBSCAN>

<https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/>

<https://towardsdatascience.com/dbscan-make-density-based-clusters-by-hand-2689dc335120>

<https://www.youtube.com/watch?v=NO8-UurDrQw>

# Reference (2)

[https://en.wikipedia.org/wiki/Hierarchical\\_clustering](https://en.wikipedia.org/wiki/Hierarchical_clustering)

<https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/>

<https://www.geeksforgeeks.org/hierarchical-clustering/>

<https://www.javatpoint.com/hierarchical-clustering-in-machine-learning>

<https://www.solver.com/xlminer/help/hierarchical-clustering-intro>

<https://builtin.com/articles/gaussian-mixture-model>

<https://medium.com/@juanc.olamendy/understanding-gaussian-mixture-models-a-comprehensive-guide-df30af59ced7>

<https://brilliant.org/wiki/gaussian-mixture-model/>

<https://www.geeksforgeeks.org/gaussian-mixture-model/>

<https://www.analyticsvidhya.com/blog/2019/10/gaussian-mixture-models-clustering/>

<https://medium.com/@rahulkaliyath/day-11-gaussian-mixture-model-clustering-90e31cdadb12>



THANK YOU