Cluster

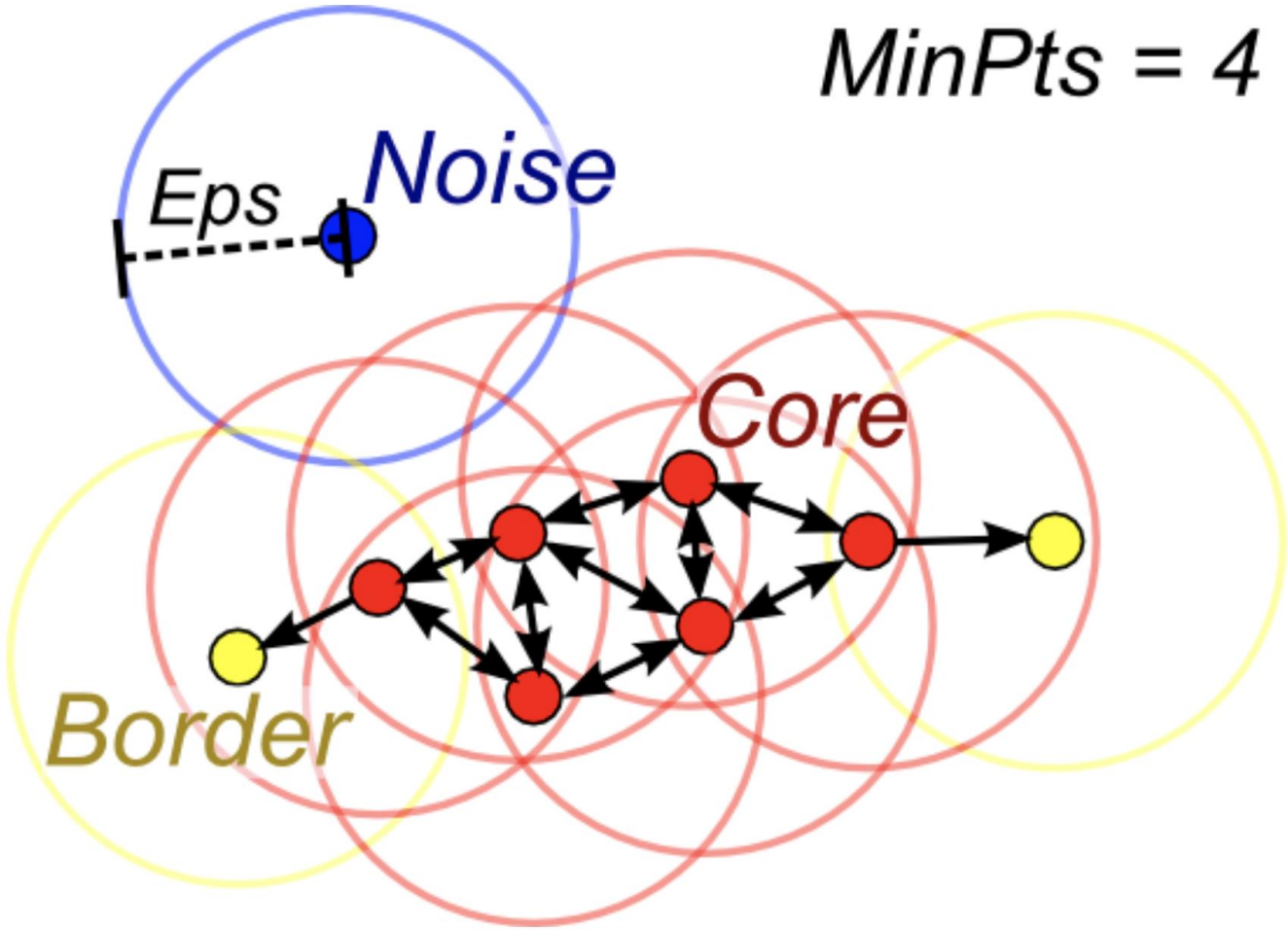
DBSCAN

- **DBSCAN** is a clustering method that is used in machine learning to separate clusters of high density from clusters of low density. Given that **DBSCAN** is a **density based clustering algorithm**, it does a great job of seeking areas in the data that have a high density of observations, versus areas of the data that are not very dense with observations. DBSCAN can sort data into clusters of varying shapes as well, another strong advantage. DBSCAN works as such:

DBSCAN

- Divides the dataset into *n* dimensions
- For each point in the dataset, DBSCAN forms an *n* dimensional shape around that data point, and then counts how many data points fall within that shape.
- DBSCAN counts this shape as a *cluster.* DBSCAN iteratively expands the cluster, by going through each individual point within the cluster, and counting the number of other data points nearby. Take the graphic below for an example:

DBSCAN

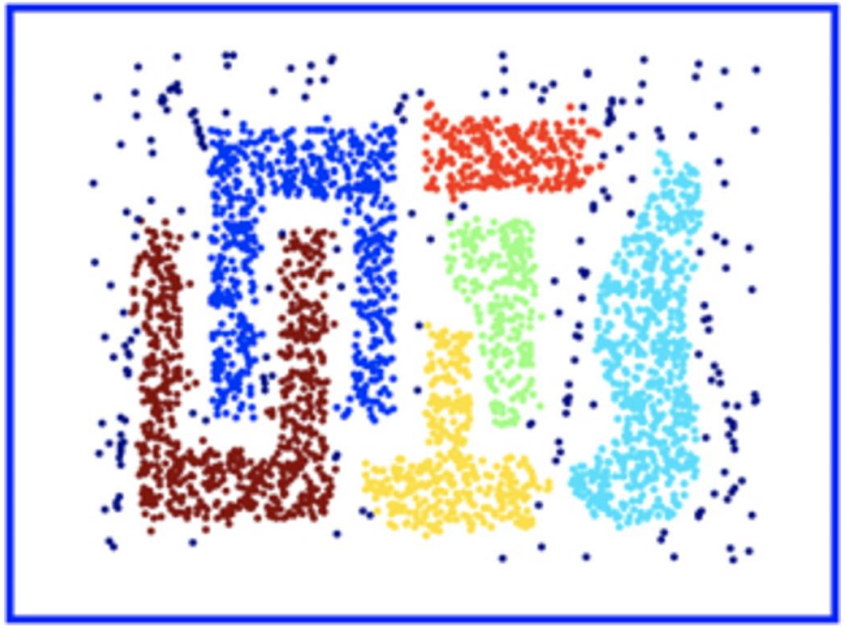


Red: Core Points

Yellow: Border points. Still part of the cluster because it's within epsilon of a core point, but not does not meet the min_points criteria

Blue: Noise point. Not assigned to a cluster

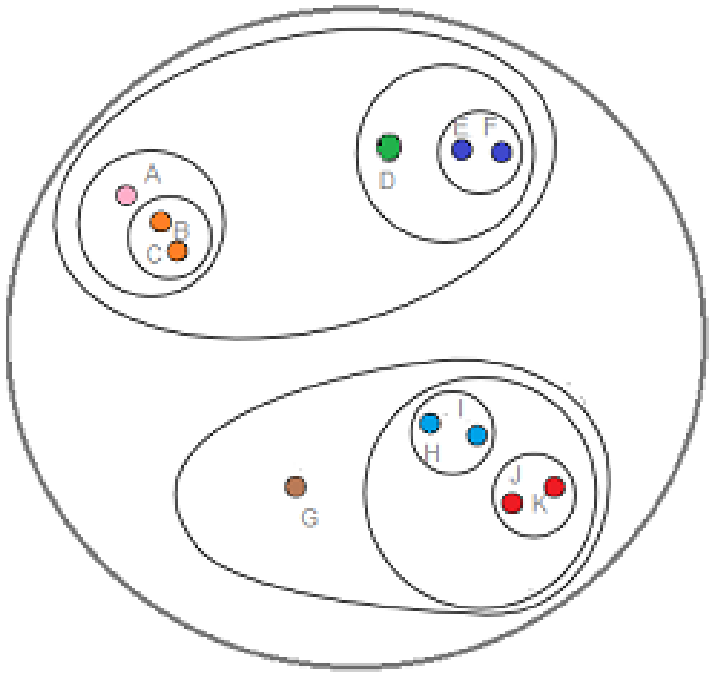
DBSCAN - Example

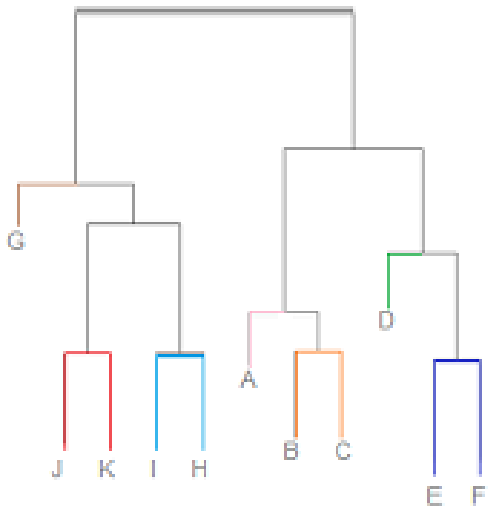


DBSCAN - Summary

- A cluster then satisfies two properties:
  - All points within the cluster are mutually density-connected.
  - If a point is density-reachable from any point of the cluster, it is part of the cluster as well.

Hierarchical Clustering





- Hierarchical Clustering - בנית קבוצות של דוגמאות וארגונם בעצים, ע"פ יחסי ההיררכיה בניהם
- כל קצה בגרף העצים מייצג חלק מהקבוצה
- הבניה של הקבוצות נעשית ע"פ המרחקים בין הדוגמאות

Hierarchical Clustering - Dendrogram