



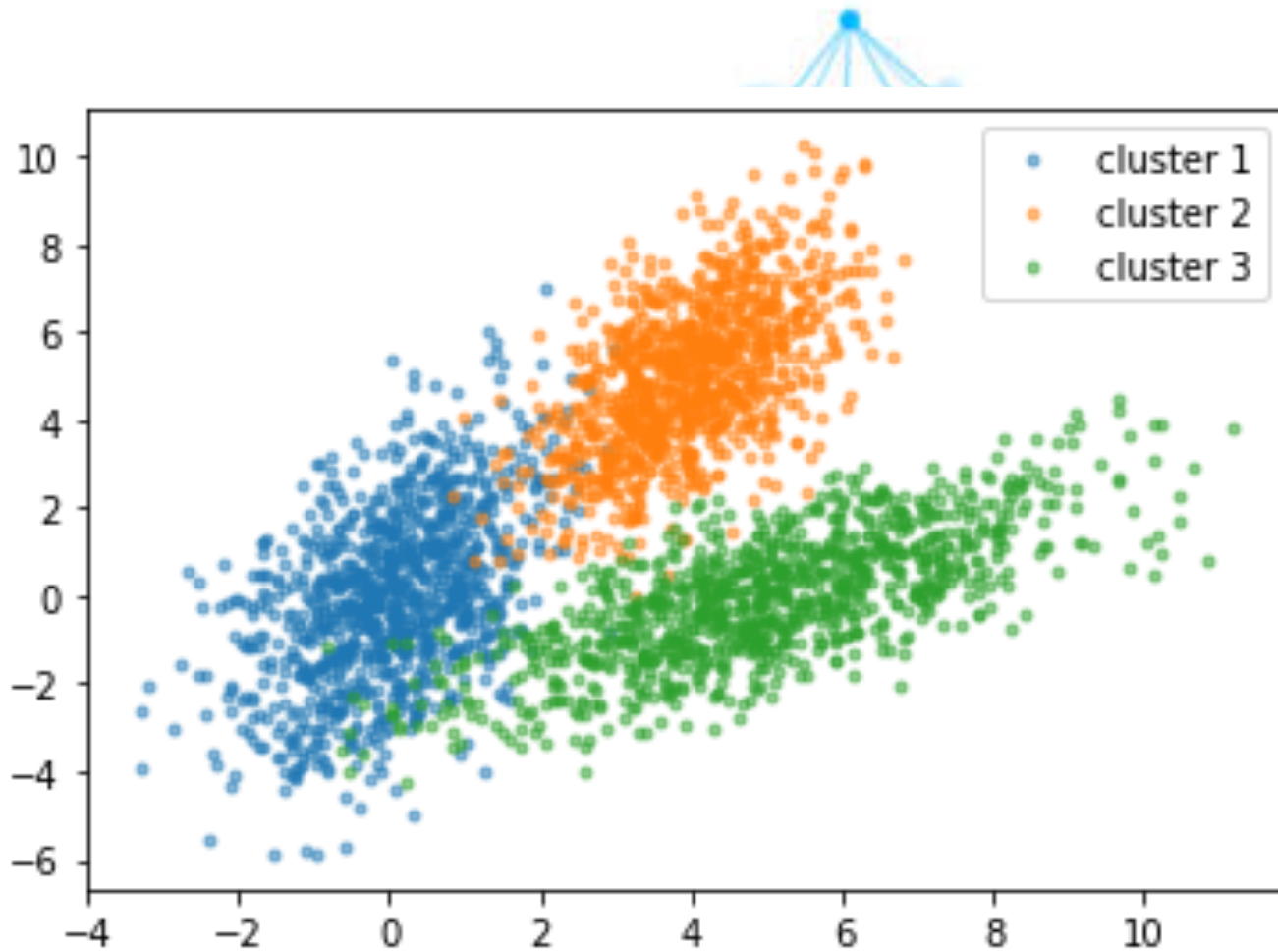
A faint, light blue network diagram serves as the background. It consists of numerous small circular nodes connected by thin, straight lines, forming a complex web of connections. The nodes are distributed across the slide, with a higher density in the upper half.

Fundamentals of Big Data Analytics

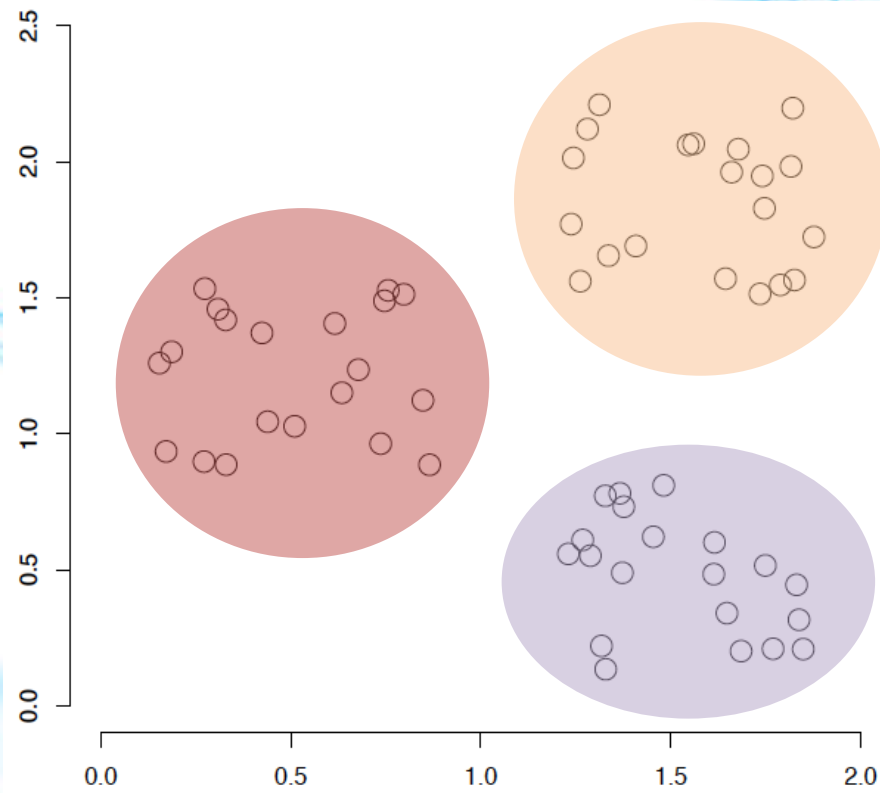
Lecture 5- Clustering & Cluster Analysis

Dr. Iqra Safder
Assistant Professor
FAST NUCES, Lahore

What is Clustering?



A data set with clear cluster structure



• 3 Clusters

A data set with clear cluster structure

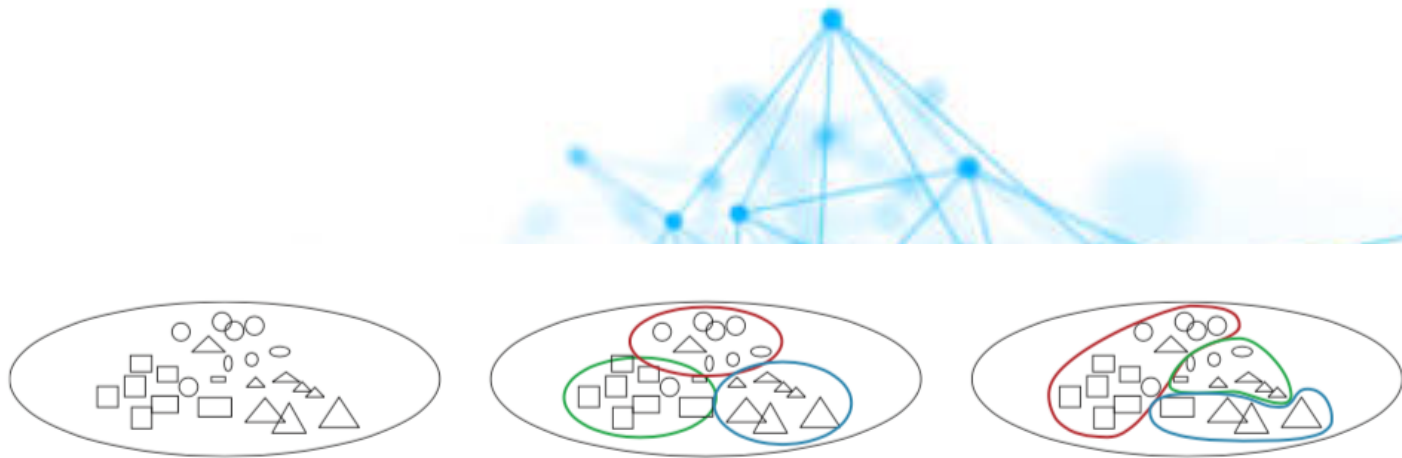
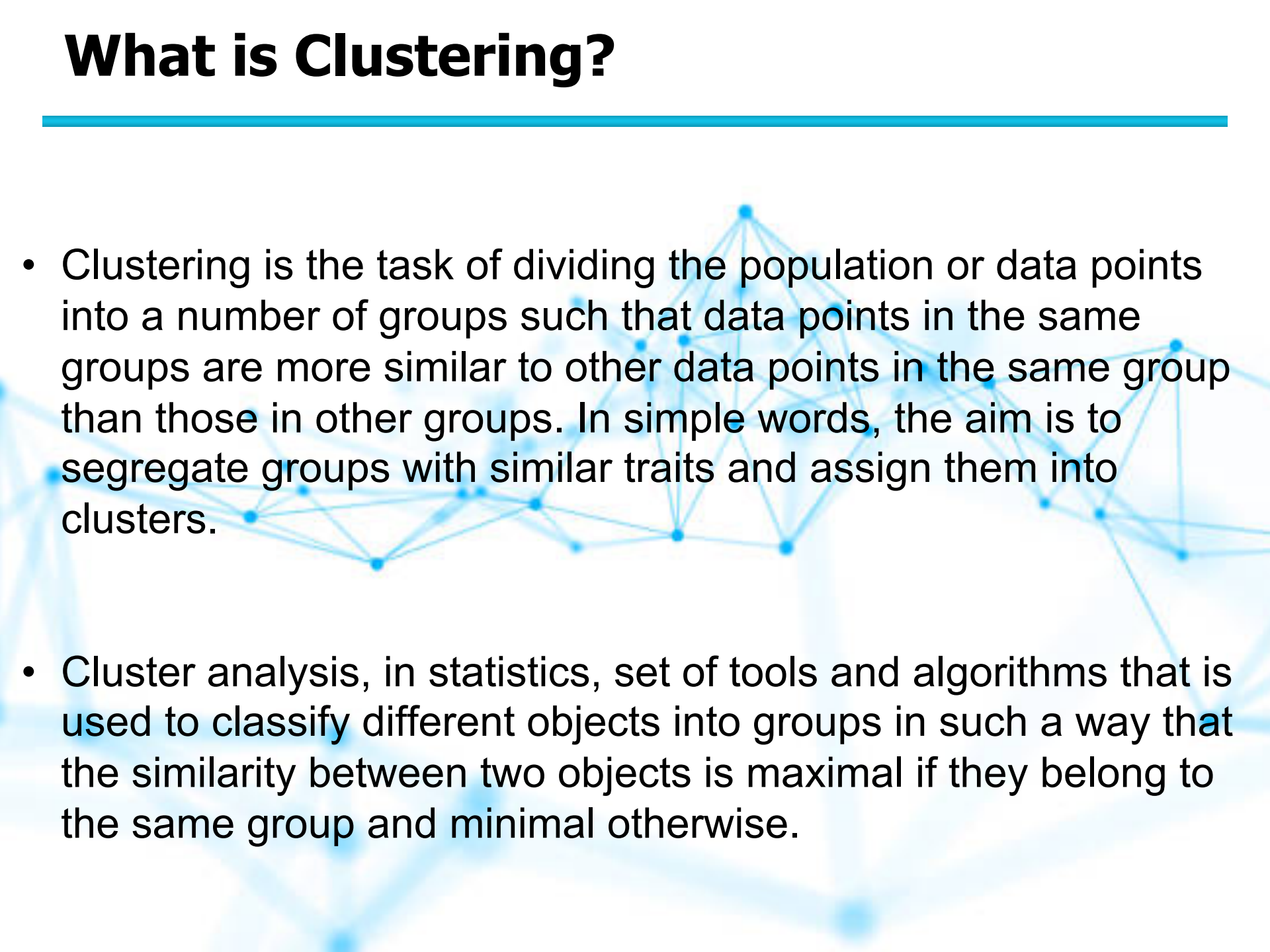


Figure 14.1 Illustration of clustering bias. The figure on the left shows a set of objects that can be potentially clustered in different ways depending on the definition of similarity (or clustering bias). The figure in the middle shows the clustering results when similarity is defined based on the shape of an object. The figure on the right shows the clustering results of the same set of objects when similarity is defined based on size.

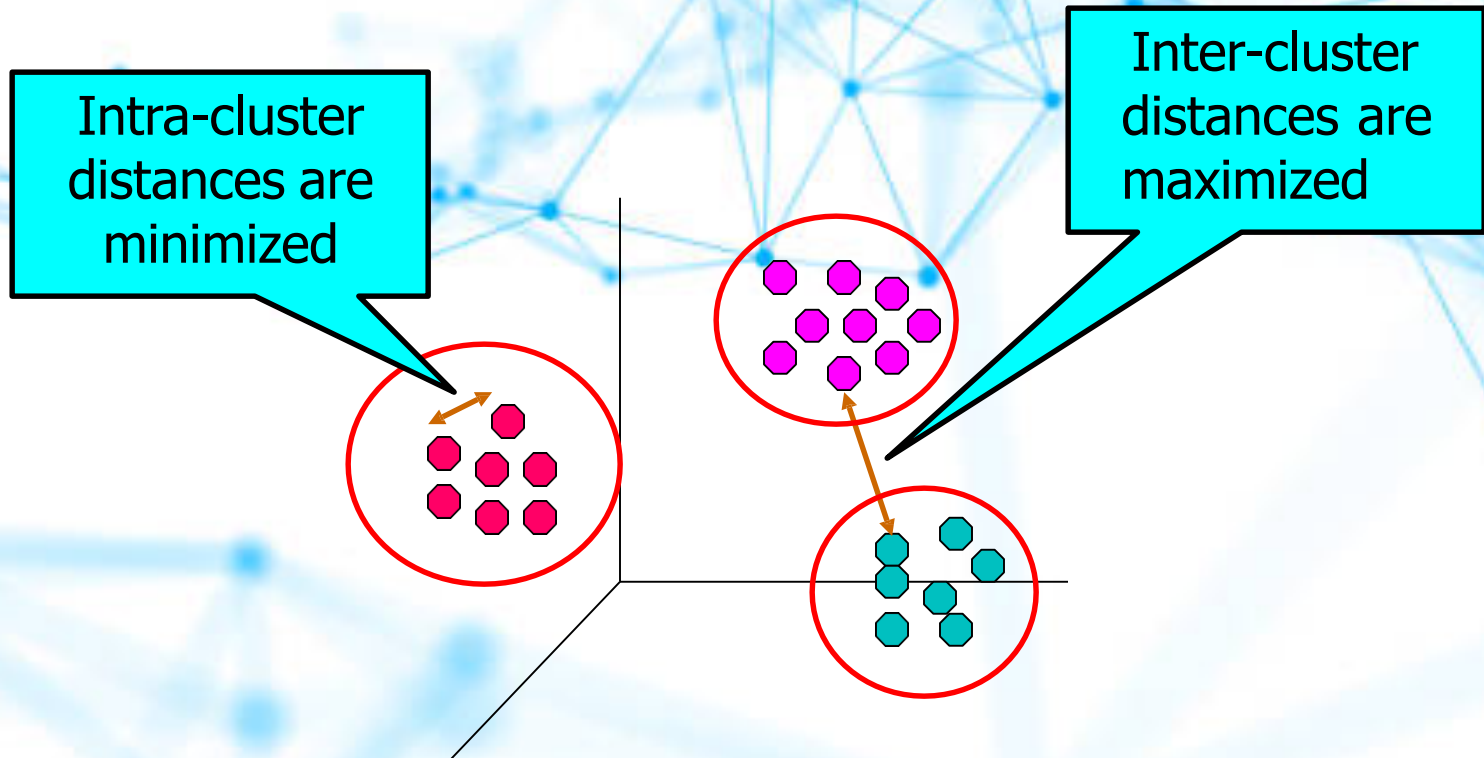
Car and Horse are similar?

What is Clustering?

- 
- Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.
 - Cluster analysis, in statistics, set of tools and algorithms that is used to classify different objects into groups in such a way that the similarity between two objects is maximal if they belong to the same group and minimal otherwise.

What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Applications of Cluster Analysis

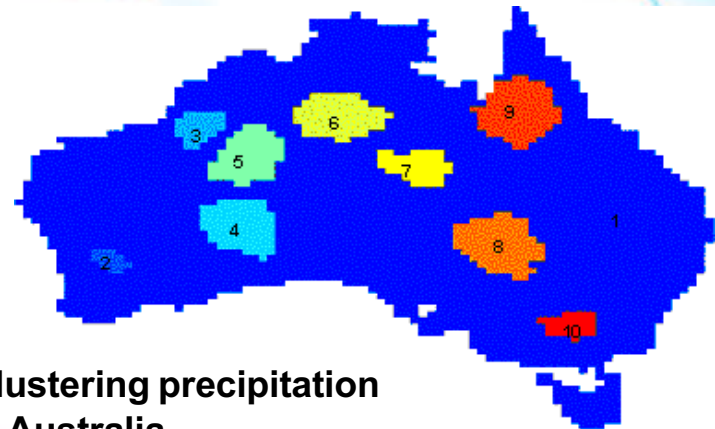
□ Understanding

- Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations

	<i>Discovered Clusters</i>	<i>Industry Group</i>
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN,Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN,DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN,Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down,Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN,Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN,ADV-Micro-Device-DOWN,Andrew-Corp-DOWN,Computer-Assoc-DOWN,Circuit-City-DOWN,Compaq-DOWN,EMC-Corp-DOWN,Gen-Inst-DOWN,Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN,MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP,Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP,Schlumberger-UP	Oil-UP

□ Summarization

- Reduce the size of large data sets



Clustering precipitation
in Australia

What is not Cluster Analysis?

- Supervised classification

- Have class label information

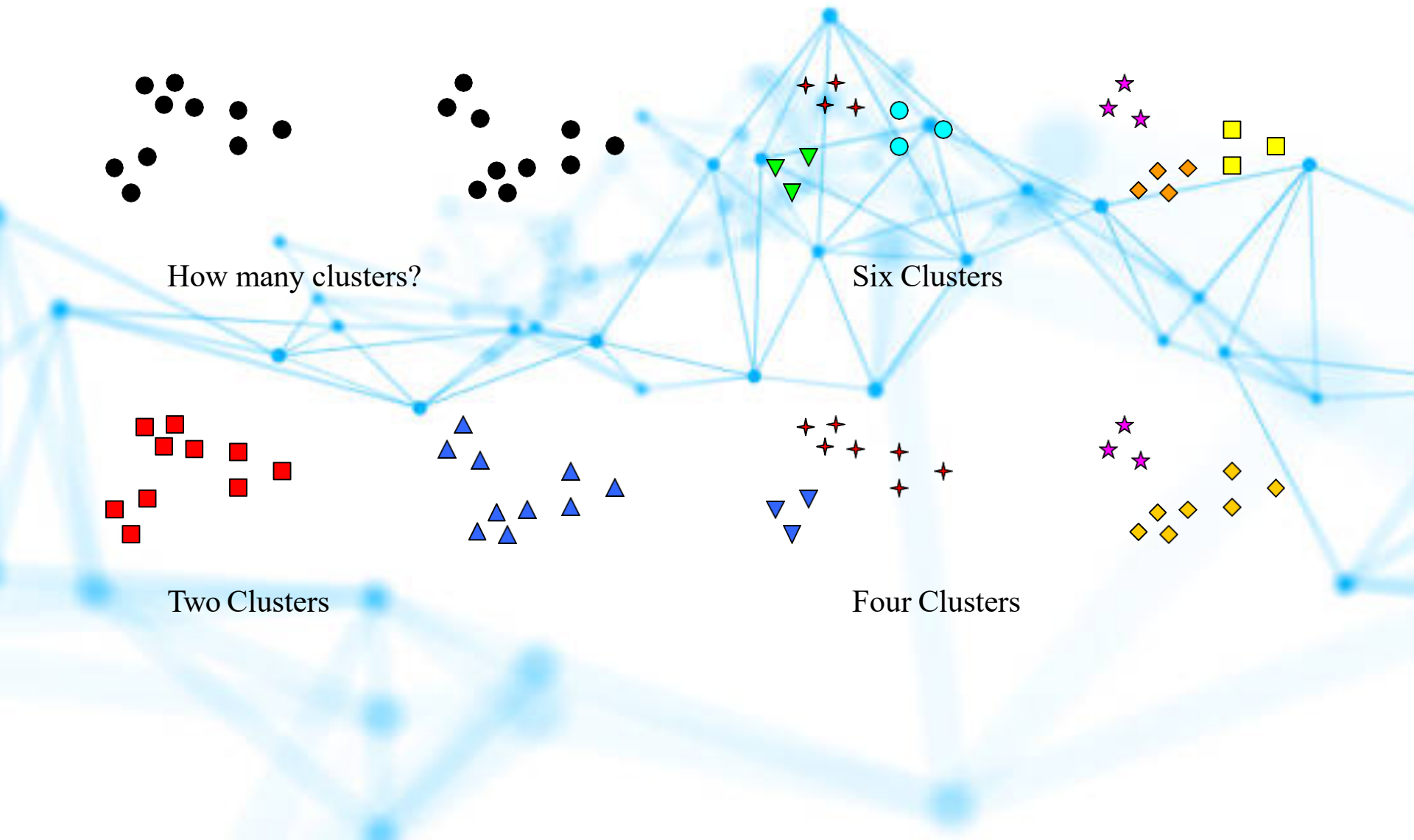
- Simple segmentation

- Dividing students into different registration groups alphabetically, by last name

- Results of a query

- Groupings are a result of an external specification

Notion of a Cluster can be Ambiguous



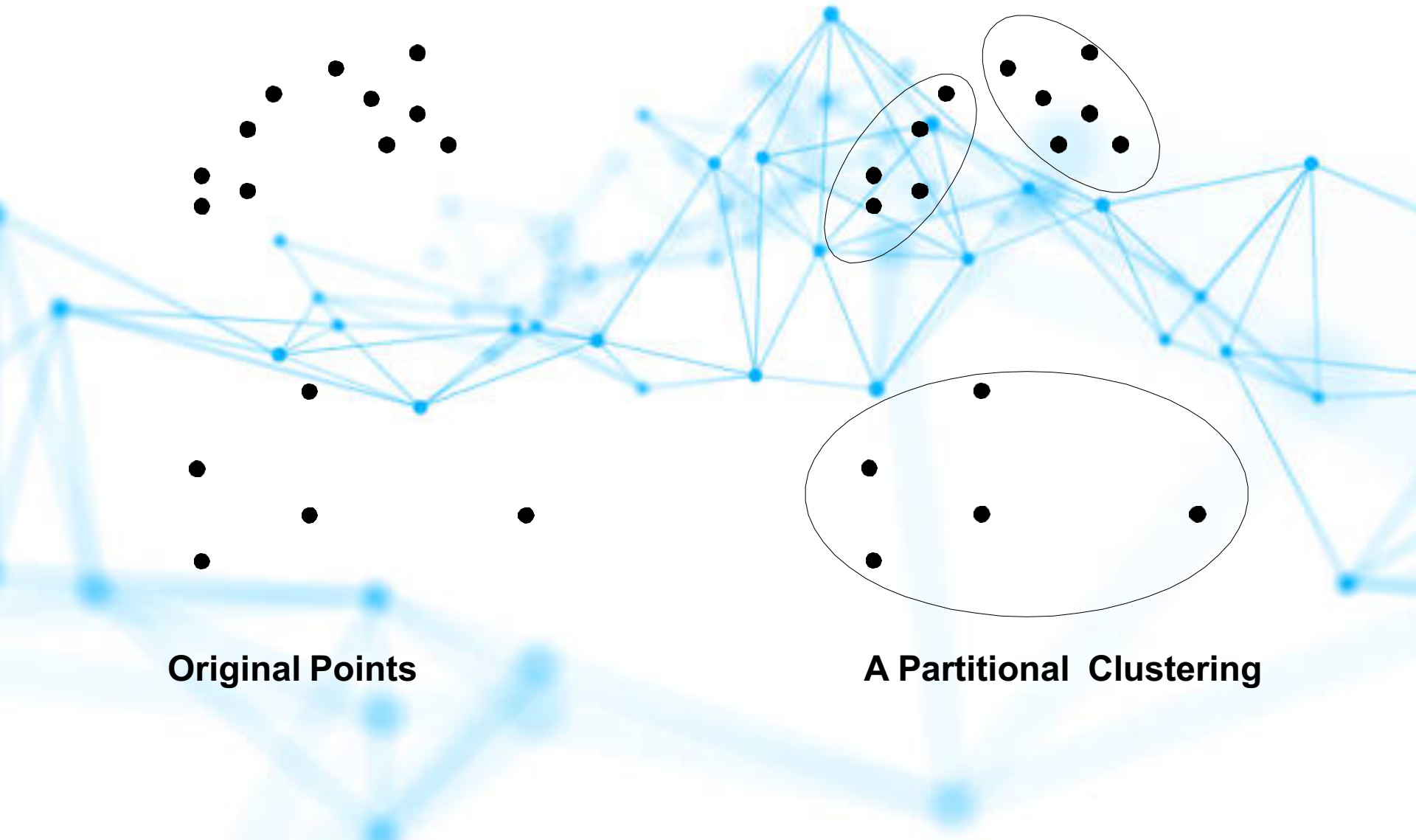
Hard vs. soft clustering

- **Hard clustering:** Each document belongs to exactly one cluster
 - More common and easier to do
- **Soft clustering:** A document can belong to more than one cluster.
 - Makes more sense for applications like creating browsable hierarchies
 - You may want to put a pair of sneakers in two clusters:
(i) sports apparel and (ii) shoes
 - You can only do that with a soft clustering approach.

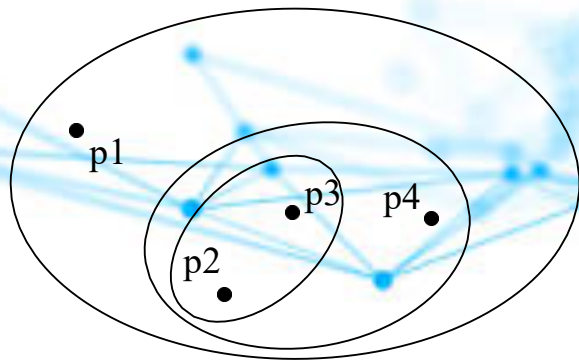
Types of Clusterings

- A **clustering** is a set of clusters
- Important distinction between **hierarchical** and **partitional** sets of clusters
- Partitional Clustering
 - A division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering
 - A set of nested clusters organized as a hierarchical tree

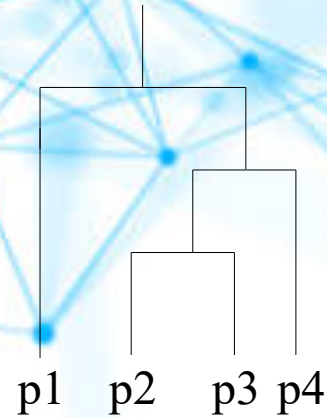
Partitional Clustering



Hierarchical Clustering



Traditional Hierarchical Clustering



Traditional Dendrogram

Other Distinctions Between Sets of Clusters

□ Exclusive versus non-exclusive

- In non-exclusive clusterings, points may belong to multiple clusters.
- Can represent multiple classes or 'border' points

□ Fuzzy versus non-fuzzy

- In fuzzy clustering, a membership point belongs to every cluster with some weight between 0 and 1
- Weights must sum to 1
- Probabilistic clustering has similar characteristics

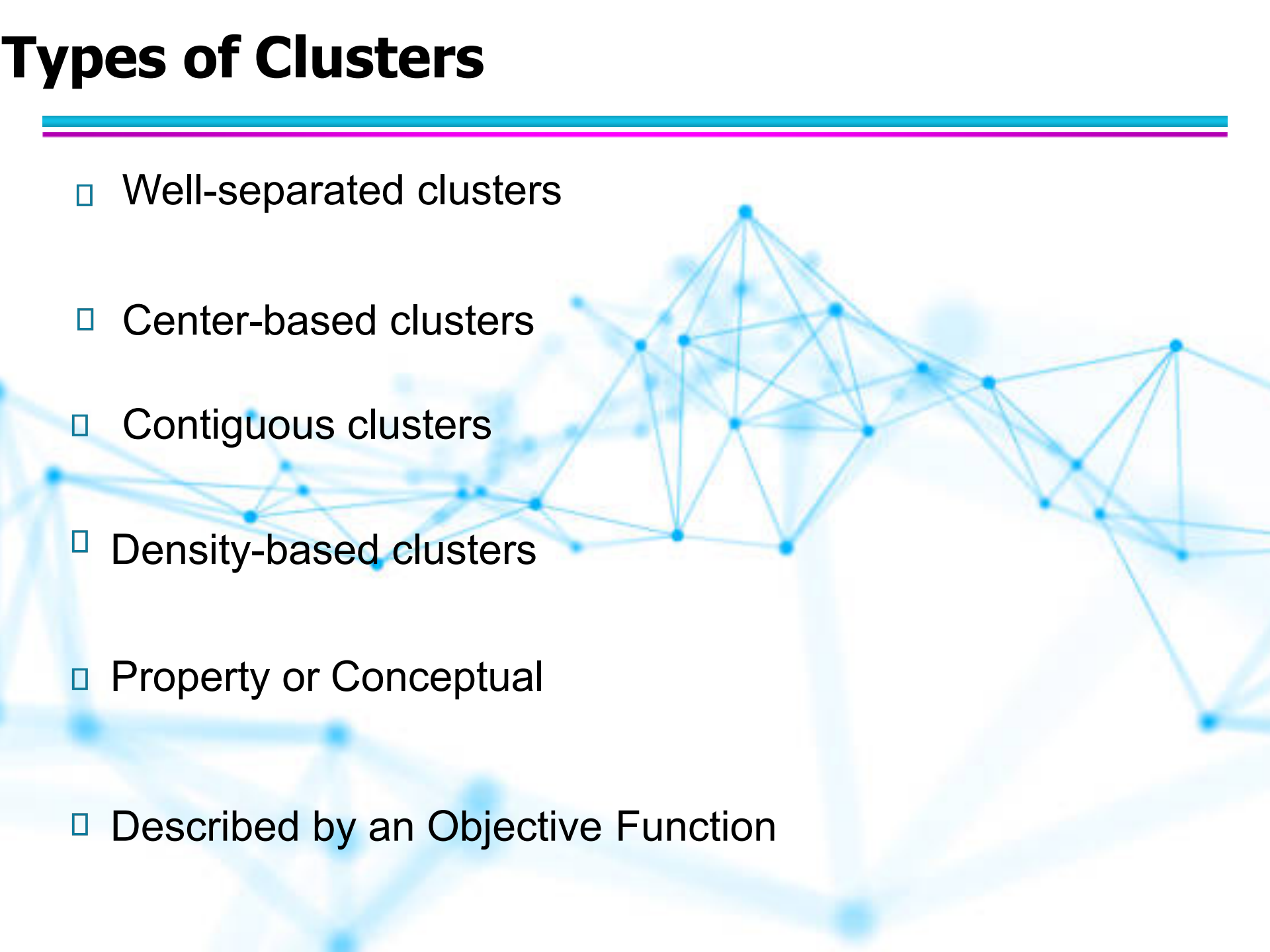
□ Partial versus complete

- In some cases, we only want to cluster some of the data

□ Heterogeneous versus homogeneous

- Cluster of widely different sizes, shapes, and densities

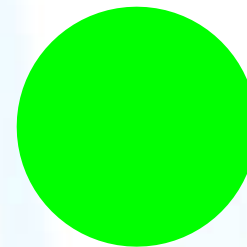
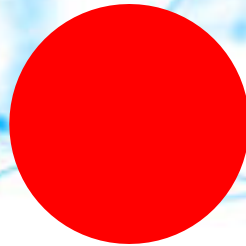
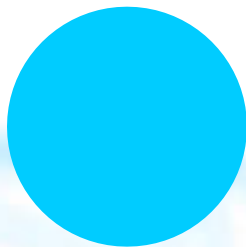
Types of Clusters

- Well-separated clusters
 - Center-based clusters
 - Contiguous clusters
 - Density-based clusters
 - Property or Conceptual
 - Described by an Objective Function
- 

Types of Clusters: Well-Separated

□ Well-Separated Clusters:

- A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.

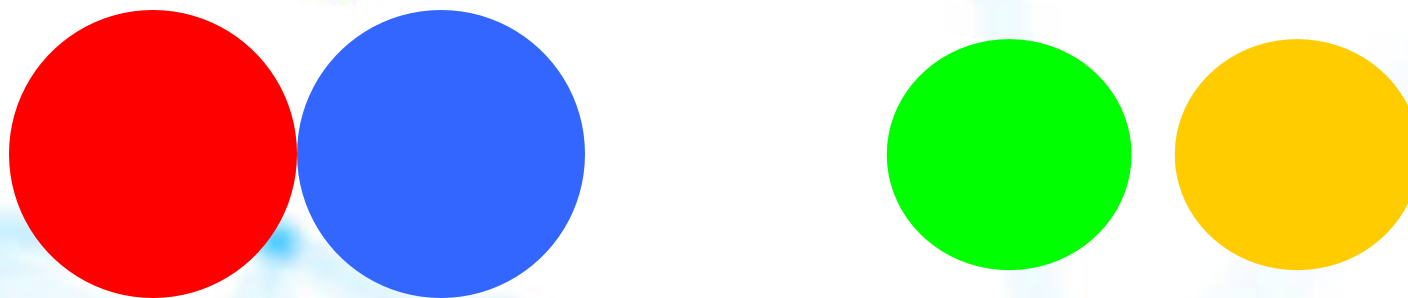


3 well-separated clusters

Types of Clusters: Center-Based

□ Center-based

- A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster
- The center of a cluster is often a **centroid**, the average of all the points in the cluster, or a **medoid**, the most “representative” point of a cluster

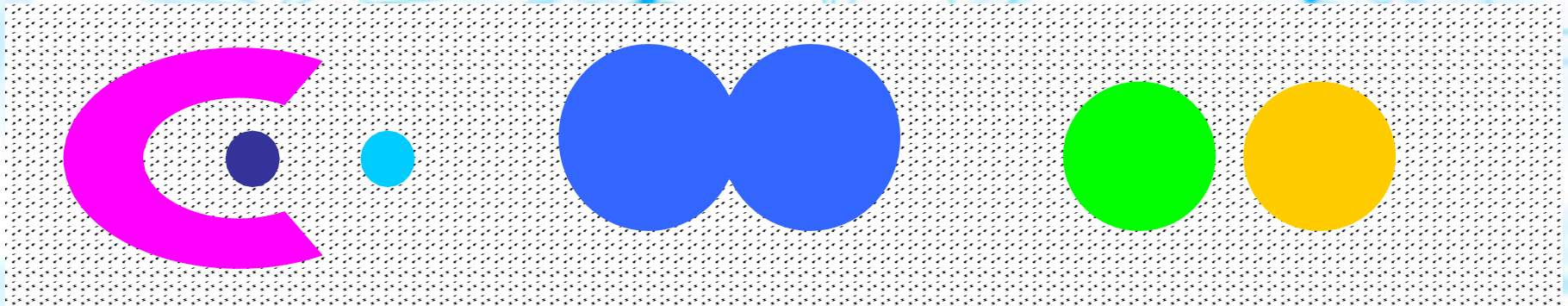


4 center-based clusters

Types of Clusters: Density-Based

□ Density-based

- A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
- Used when the clusters are irregular or intertwined, and when noise and outliers are present.



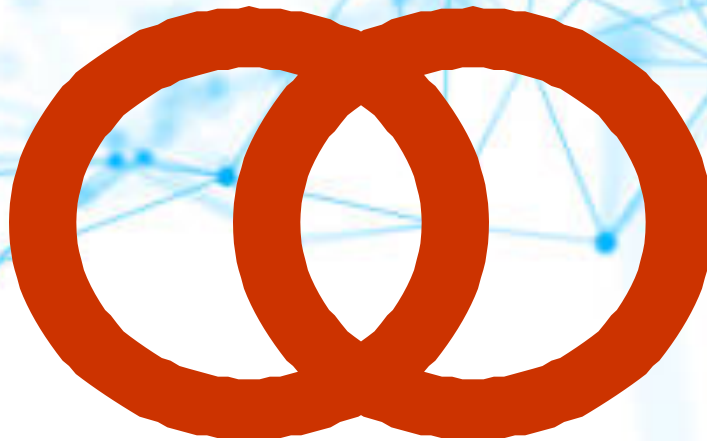
6 density-based clusters

Types of Clusters: Conceptual Clusters

□ Shared Property or Conceptual Clusters

- Finds clusters that share some common property or represent a particular concept.

.



2 Overlapping Circles

Types of Clusters: Objective Function

□ Clusters Defined by an Objective Function

- Finds clusters that minimize or maximize an objective function.
- Enumerate all possible ways of dividing the points into clusters and evaluate the 'goodness' of each potential set of clusters by using the given objective function. (NP Hard)
- Can have global or local objectives.
 - ◆ Hierarchical clustering algorithms typically have local objectives
 - ◆ Partitional algorithms typically have global objectives