

The background of the slide is a complex, abstract composition. It features a central white banner with a subtle, light gray geometric pattern. Surrounding this banner are various abstract elements: a network of thin, reddish-brown lines forming a web-like structure, interspersed with small green dots; a grid of small, light gray plus signs; and a series of vertical, slightly wavy lines in shades of brown and orange. In the top left corner, there is a small, rectangular inset image showing a cluster of orange and red dots on a light background, with a faint grid pattern overlaid.

Basic Concepts: Measuring Similarity between Objects

What Is Good Clustering?

- A good clustering method will produce high quality clusters which should have
 - **High intra-class similarity:** **Cohesive** within clusters 类内相似度高.
 - **Low inter-class similarity:** **Distinctive** between clusters 类间相似度低.
- **Quality function**
 - There is usually a separate “quality” function that measures the “goodness” of a cluster
 - It is hard to define “similar enough” or “good enough”
 - The answer is typically highly subjective
- There exist many similarity measures and/or functions for different applications
- Similarity measure is critical for cluster analysis

o

Similarity, Dissimilarity, and Proximity

□ Similarity measure or similarity function

- A real-valued function that quantifies the similarity between two objects
- Measure how two data objects are alike: The higher value, the more alike
- Often falls in the range $[0,1]$: 0: no similarity; 1: completely similar

□ Dissimilarity (or distance) measure

- Numerical measure of how different two data objects are
- In some sense, the inverse of similarity: The lower, the more alike
- Minimum dissimilarity is often 0 (i.e., completely similar)
- Range $[0, 1]$ or $[0, \infty)$, depending on the definition

□ Proximity usually refers to either similarity or dissimilarity