

The background of the slide is a complex, abstract composition. It features a dark, muted red or maroon color palette. Overlaid on this are various geometric patterns, including a network of thin, light-colored lines forming a mesh or web-like structure. There are also clusters of small, colored dots (green, blue, orange) scattered across the background. In the upper left corner, there is a small, semi-transparent inset showing a grid of data points with some points highlighted in orange and red. The title text is centered in a white, bold, sans-serif font.

# **An Overview of Typical Clustering Methodologies**



# Typical Clustering Methodologies (I)

## □ Distance-based methods

- Partitioning algorithms: K-Means, K-Medians, K-Medoids (分割)
- Hierarchical algorithms: Agglomerative vs. divisive methods (分級)

## □ Density-based and grid-based methods

down→top      top→down

- Density-based: Data space is explored at a high-level of granularity and then post-processing to put together dense regions into an arbitrary shape
- Grid-based: Individual regions of the data space are formed into a grid-like structure

## □ Probabilistic and generative models: Modeling data from a generative process

- Assume a specific form of the generative model (e.g., mixture of Gaussians)
- Model parameters are estimated with the Expectation-Maximization (EM) algorithm (using the available dataset, for a maximum likelihood fit)
- Then estimate the generative probability of the underlying data points



# Typical Clustering Methodologies (II)

## ■ High-dimensional clustering

- Subspace clustering: Find clusters on various subspaces
  - Bottom-up, top-down, correlation-based methods vs.  $\delta$ -cluster methods
- Dimensionality reduction: A vertical form (i.e., columns) of clustering
  - Columns are clustered; may cluster rows and columns together (co-clustering)
- Probabilistic latent semantic indexing (PLSI) then LDA: Topic modeling of text data
  - A cluster (i.e., topic) is associated with a set of words (i.e., dimensions) and a set of documents (i.e., rows) simultaneously
- Nonnegative matrix factorization (NMF) (as one kind of co-clustering)
  - A nonnegative matrix  $A$  (e.g., word frequencies in documents) can be approximately factorized two non-negative low rank matrices  $U$  and  $V$
- Spectral clustering: Use the *spectrum* of the similarity matrix of the data to perform dimensionality reduction for clustering in fewer dimensions