

The background features a complex, abstract design. It includes a network of thin, reddish-brown lines connecting various points, some of which are green dots. There are also larger, fainter geometric shapes and patterns in shades of brown and grey. A prominent white, angular shape, resembling a stylized 'V' or a folded piece of paper, is positioned behind the title text. On the left side, there is a small inset image showing a cluster of orange and red dots on a light background, with a grid of small '+' markers nearby.

# The *K-Medians* and *K-Modes* Clustering Methods



# ***K-Medians*: Handling Outliers by Computing Medians**

- Medians are less sensitive to outliers than means
  - Think of the median salary vs. mean salary of a large firm when adding a few top executives!
- ***K-Medians***: Instead of taking the **mean** value of the object in a cluster as a reference point, **medians** are used ( $L_1$ -norm as the distance measure)
- The criterion function for the *K-Medians* algorithm: 
$$S = \sum_{k=1}^K \sum_{x_{ij} \in C_k} |x_{ij} - med_{kj}|$$
- The *K-Medians* clustering algorithm:
  - Select  $K$  points as the initial representative objects (i.e., as initial  $K$  medians)
  - **Repeat**
    - Assign every point to its nearest median
    - Re-compute the median using the median of each individual feature
  - **Until** convergence criterion is satisfied

# K-Modes: Clustering Categorical Data

- ❑ *K-Means* cannot handle non-numerical (categorical) data
  - ❑ Mapping categorical value to 1/0 cannot generate quality clusters for high-dimensional data
- ❑ **K-Modes**: An extension to *K-Means* by replacing means of clusters with **modes**
- ❑ Dissimilarity measure between object  $X$  and the center of a cluster  $Z$ 
  - ❑  $\Phi(x_j, z_j) = 1 - n_j^r/n_l$  when  $x_j = z_j$  ; 1 when  $x_j \neq z_j$ 
    - ❑ where  $z_j$  is the categorical value of attribute  $j$  in  $Z$ ,  $n_l$  is the number of objects in cluster  $l$ , and  $n_j^r$  is the number of objects whose attribute value is  $r$
- ❑ This dissimilarity measure (distance function) is **frequency-based**
- ❑ Algorithm is still based on iterative *object cluster assignment* and *centroid update*
- ❑ A **fuzzy K-Modes** method is proposed to calculate a **fuzzy cluster membership value** for each object to each cluster
- ❑ A mixture of categorical and numerical data: Using a **K-Prototype** method