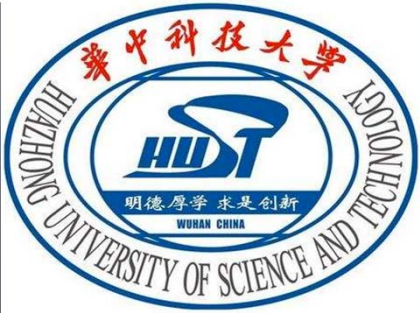
Chapter 4:
Clustering

崔金华
邮箱：jhcui@hust.edu.cn
主页: https://csjhcui.github.io/
办公地址: 东湖广场柏景阁1单元1568 室



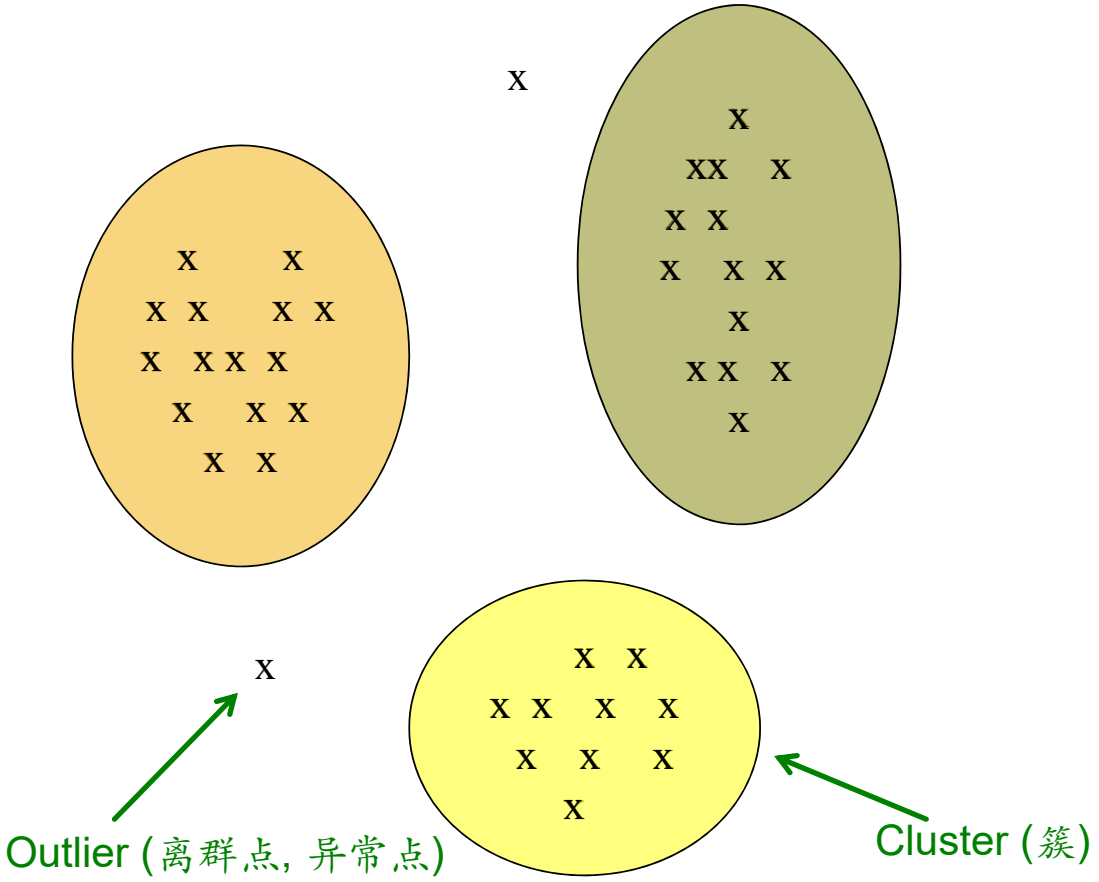
High Dimensional Data

- Given a cloud of data points we want to understand its struct

The Problem of Clustering

- Given a **set of points**, with a notion of **distance** between points, **group the points** into some number of ***clusters***, so that
  - Members of a cluster are close/similar to each other
  - Members of different clusters are dissimilar
- **Usually:**
  - Points are in a high-dimensional space
  - Similarity is defined using a distance measure
    - Euclidean, Cosine, Jaccard, edit distance, …

Example: Clusters & Outliers



Clustering is a hard problem!