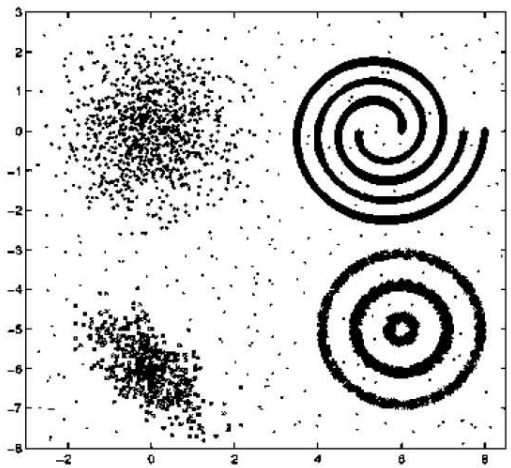
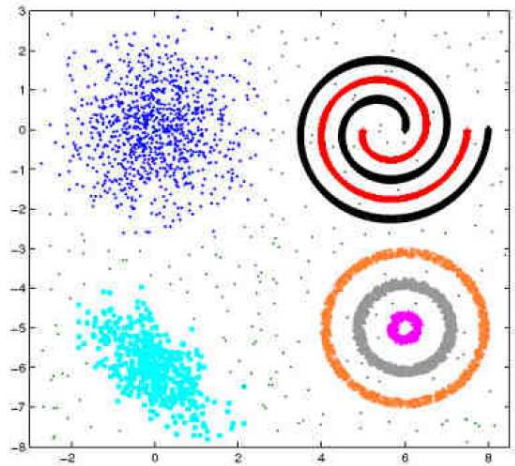
What is Data Clustering?

- Data Clustering is an unsupervised learning problem
- Given: N unlabeled examples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$; the number of partitions K
- Goal: Group the examples into K partitions



(a) Input data



(b) Desired clustering

- The only information clustering uses is the similarity between examples
- Clustering groups examples based of their mutual similarities
- A good clustering is one that achieves:
  - High within-cluster similarity
  - Low inter-cluster similarity

Picture courtesy: "Data Clustering: 50 Years Beyond K-Means", A.K. Jain (2008)