Clustering

Clustering is the problem of grouping points by similarity.

Often elements come from a small number of “sources” or “explanations”, and clustering is a good way to reveal these origins.

Similarity is defined by some underlying distance function/metric.

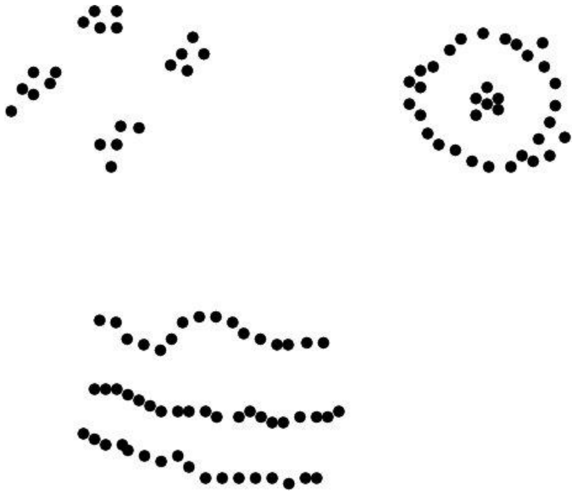
How Many Clusters Do You See?

Clustering is an inherently ill-defined problem since they upon context and the eye of the beholder.

How many do you see?

Compact, circular clusters

are natural but not universal.



Why Clustering?

- Hypothesis development -- how many distinct populations are there in your data?
- Modeling over smaller groups -- build separate predictive models for each cluster.
- Data reduction -- replace/represent each cluster of items by its centroid.
- Outlier detection -- which items are far from cluster centers, or stuck in tiny clusters?