Ciro Donalek (Caltech)

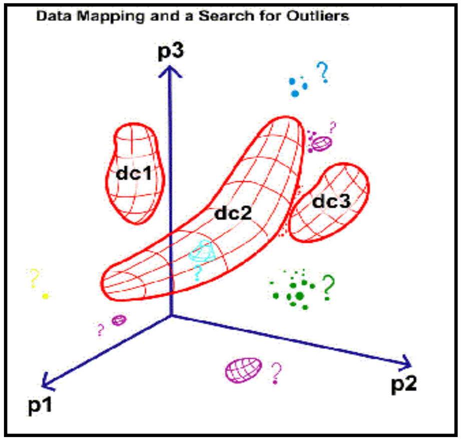
Clustering

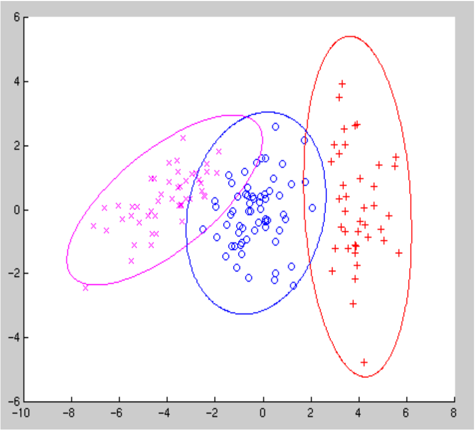
Jet Propulsion Laboratory
Keck Institute for Space Studies
Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology
Caltech

What is Clustering?

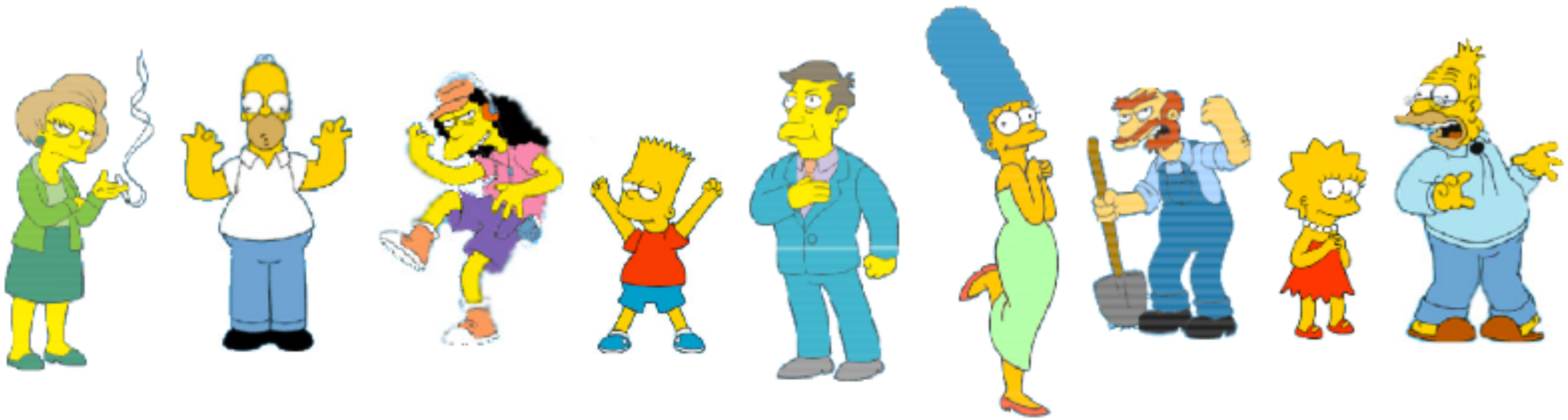
- Unsupervised Learning
- **Cluster hypothesis**: *objects in the same cluster behave similarly with respect to relevance to information needs*.
  - points in the same cluster are likely to be of the same type.
  - finding natural groupings among objects.

Unsupervised Algorithms
K-Means
Self-Organizing Maps
RDF
Fuzzy Clustering
CURE
ROCK
Vector Quantization
Probabilistic Principal Surfaces
...



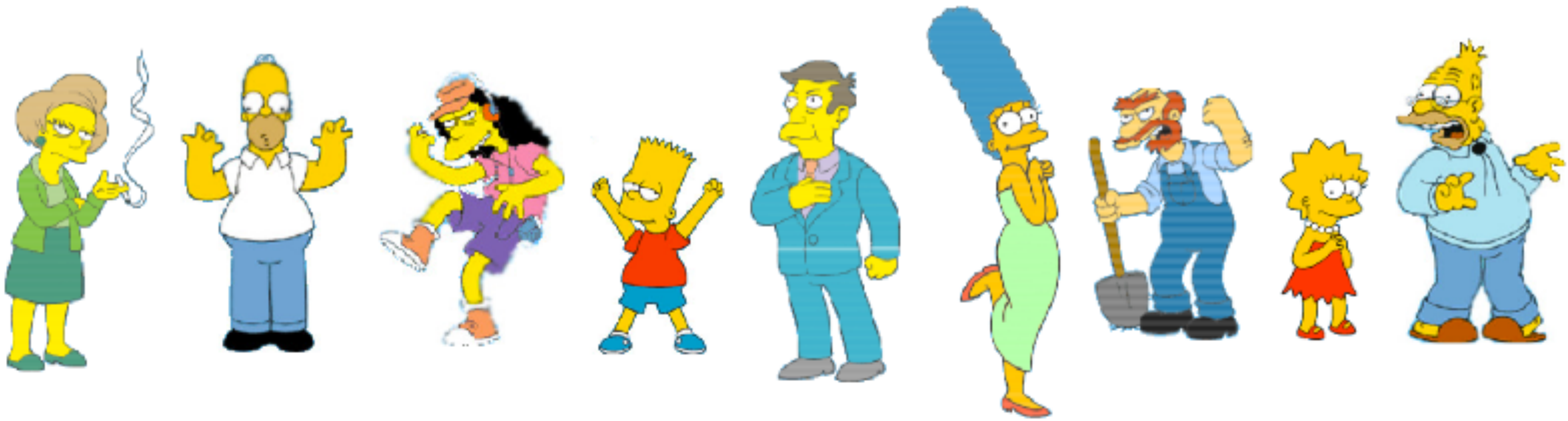


What is a natural grouping?

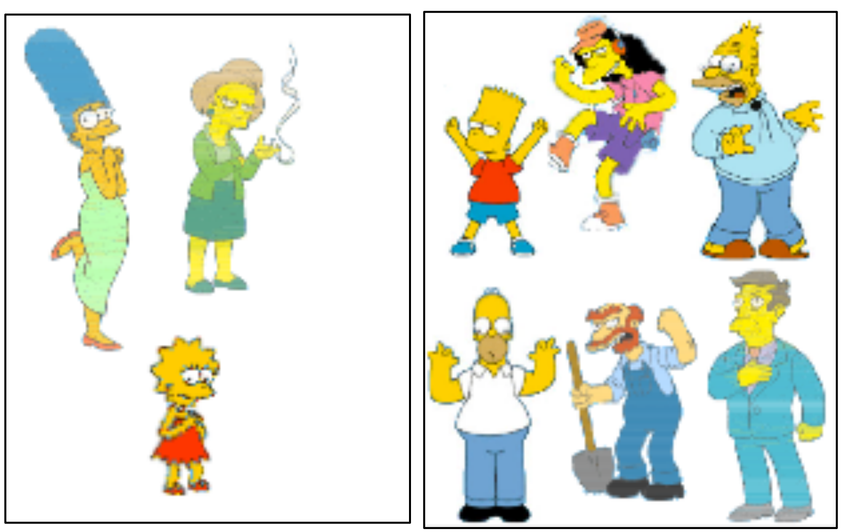


Simpson's examples from
Artificial Intelligence, Henry Lin

Clustering is subjective!



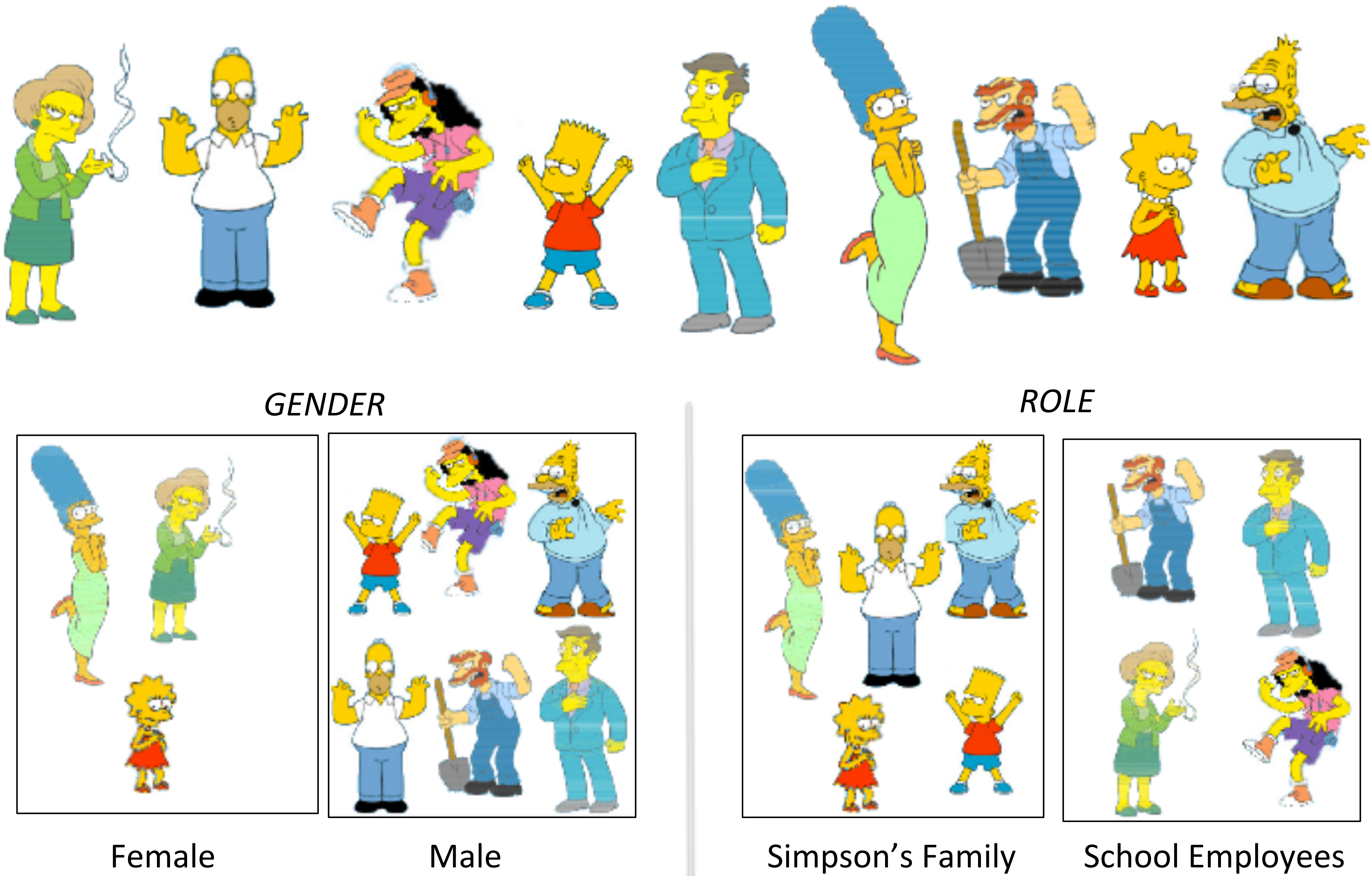
GENDER



Female

Male

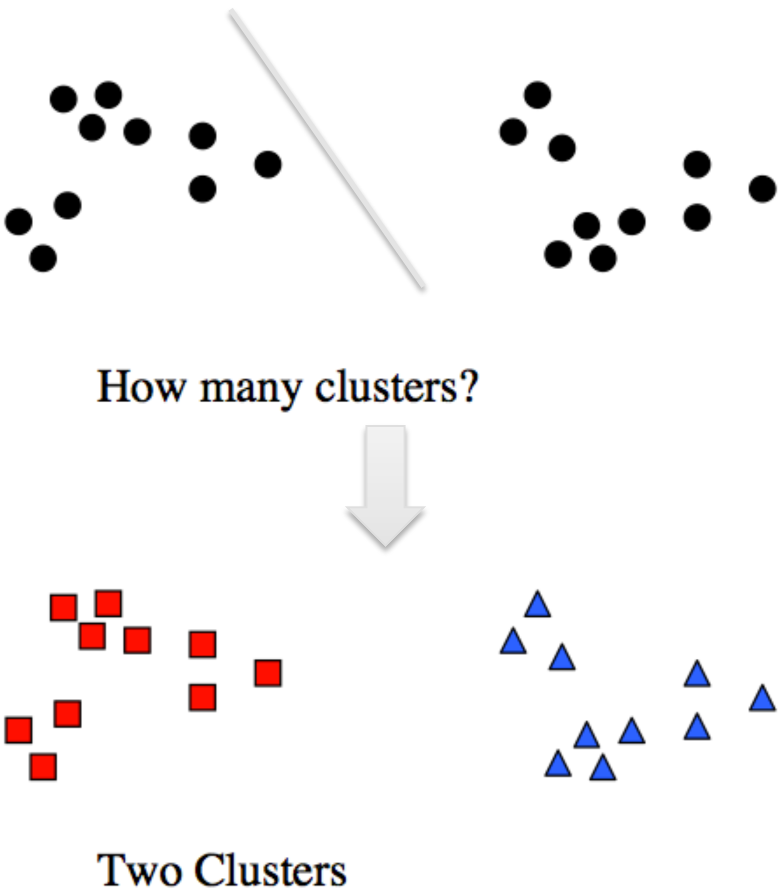
Clustering is subjective!



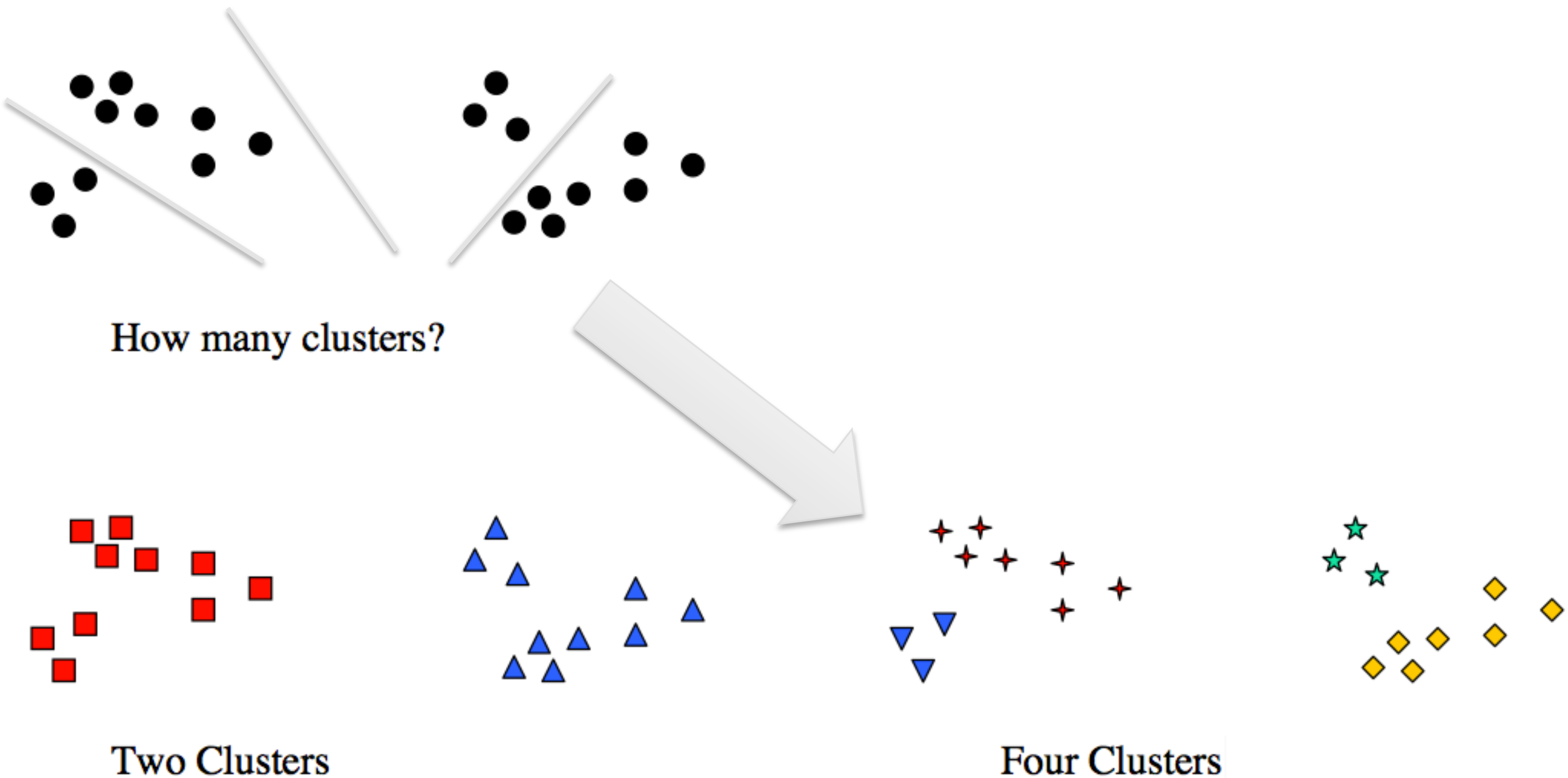
How many clusters do you see?

How many clusters?

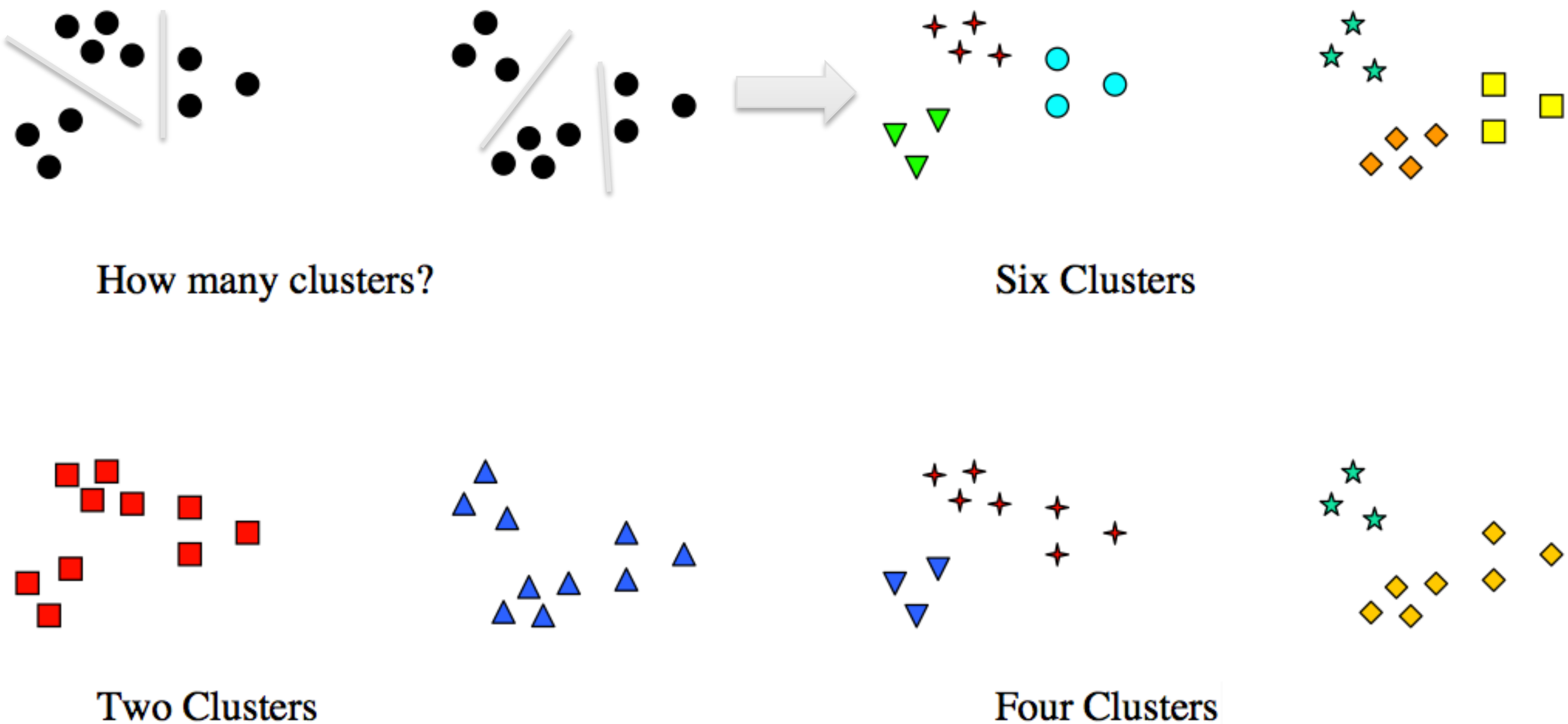
How many clusters do you see?



How many clusters do you see?

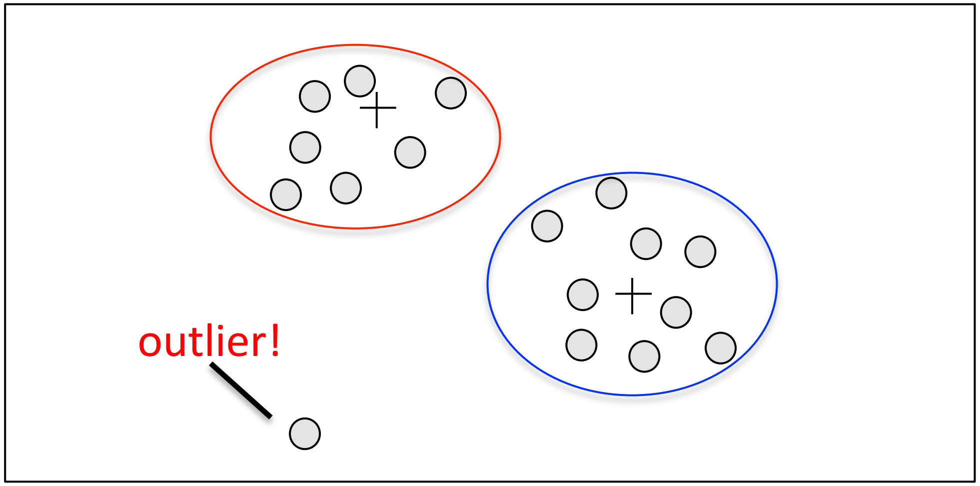


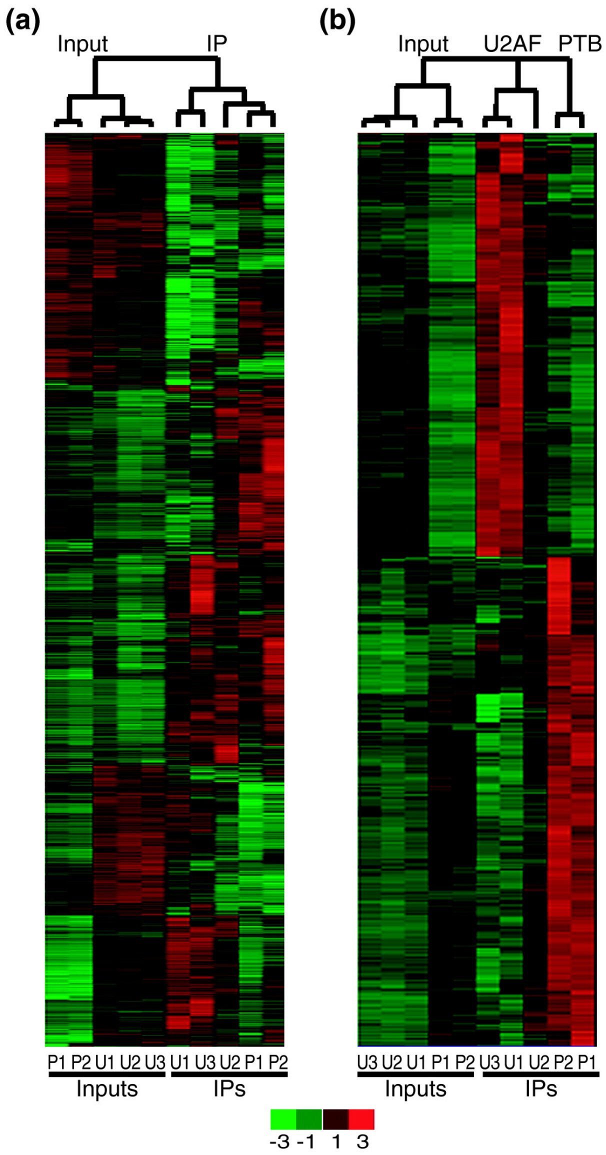
How many clusters do you see?



Why clustering

- Organizing data into clusters shows internal structure of the data.
  - e.g., gene clustering
- Use partitions to achieve a goal
  - e.g., market segmentation
- Outlier analysis, image analysis, etc.





Formal statement

- Given a set of feature vectors D:
  $D = [\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_n]$,
  a desired number of cluster k,
  an objective function g, we want to compute an assignment
  f : D -> {1,..,k} that minimize (maximize) g

- The objective function is often defined in terms of similarity or distance between samples or clusters.

Evaluation of Clustering

- Internal criterion
  - based on intra-cluster and inter-cluster similarities.
- External criterion
  - direct evaluation: may be expensive
  - benchmark or “gold standard”

Internal Measures

- A good clustering is one where:
  - the intra-cluster distance is minimized: defined as the sum of distances between objects in the same clusters;
  - the inter-cluster distance is maximized: defined as the distances between different clusters.

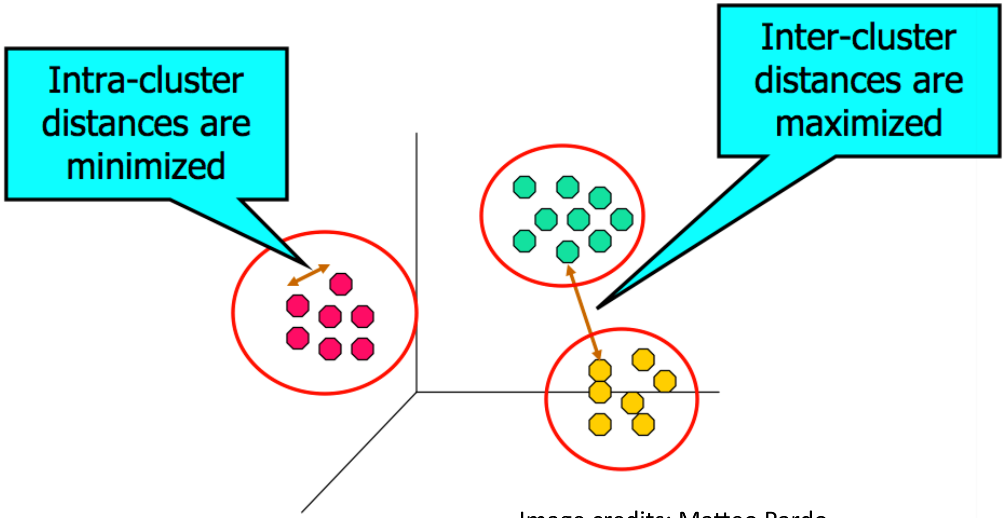


Image credits: Matteo Pardo

Internal Measures

- Good scores do not necessarily translate in good effectiveness in an application.
- Can be biased towards algorithms that use the same cluster models.
- Useful to get insights.

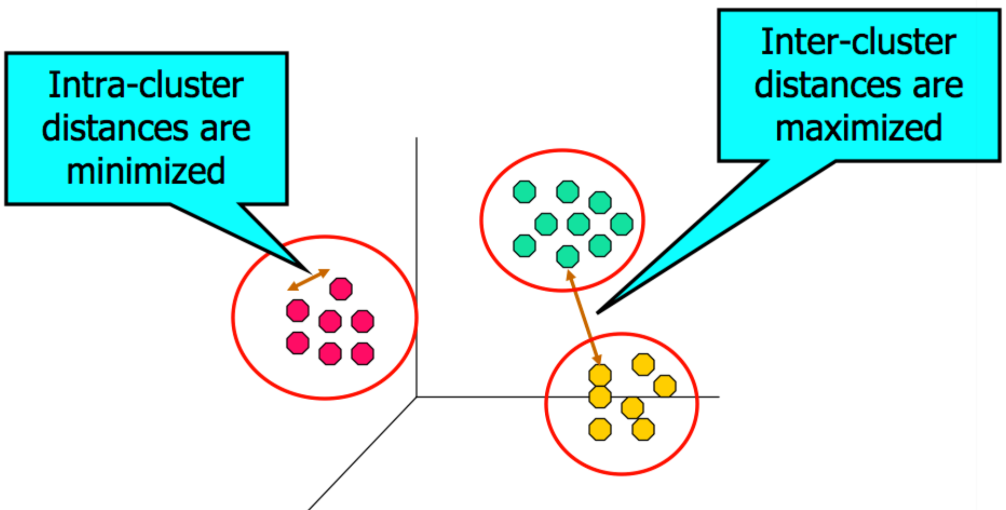


Image credits: Matteo Pardo

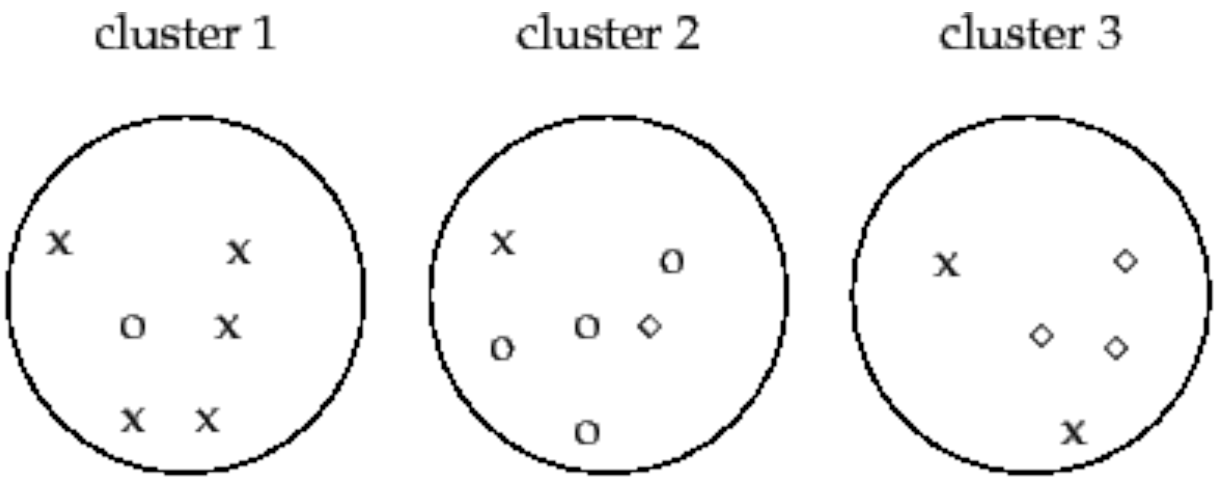
External Measures

- Use of a labeled sub-sample as gold standard.
- Evaluate how well the clustering matches the classes.
- Criteria:
  - purity;
  - normalized mutual information;
  - rand index;
  - F measure;
  - Jaccard measures.

Purity

- ***Purity***:
  - assign each cluster to the class which is most frequent;
  - measure the accuracy by counting the number of correctly assigned samples per class.
  - problem: purity=1 if each cluster contain just one sample!

Purity



► **Figure 16.1** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ⋄, 3 (cluster 3). Purity is $(1/17) \times (5 + 4 + 3) \approx 0.71$.

Ref.: Introduction to Information Retrieval
Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze
2008, Cambridge University Press

NMI

- ***Normalized Mutual Information* (NMI)**
  - measures the amount of information by which our knowledge about the classes increases when we are told what the clusters are.
  - 0 if the clustering is random with respect to the class membership.
  - maximum reached for a cluster that perfectly recreates the class, but also for a high number of clusters.

Rand Index and F measure

- The ***Rand index*** **( )** measures the percentage of decisions that are correct (accuracy).

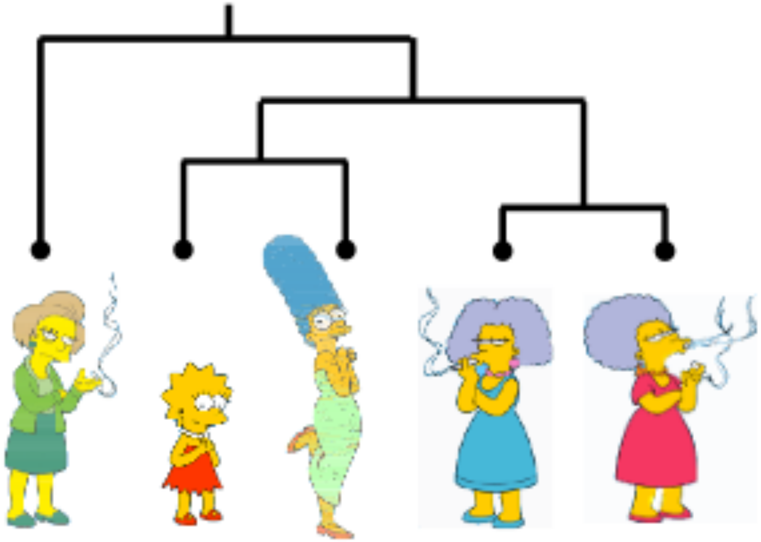
$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

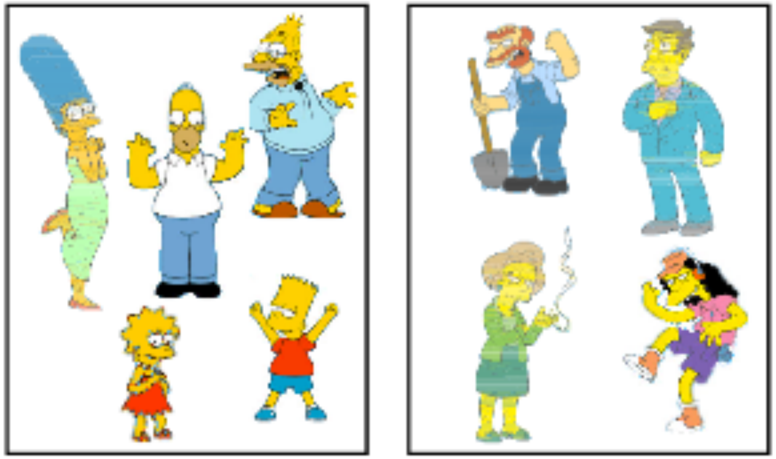
- Gives equal weight to false positives and false negatives.
- ***F measure***: introduce weights.
  - e.g., penalizes false negatives, selecting beta > 1

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Types of Clustering

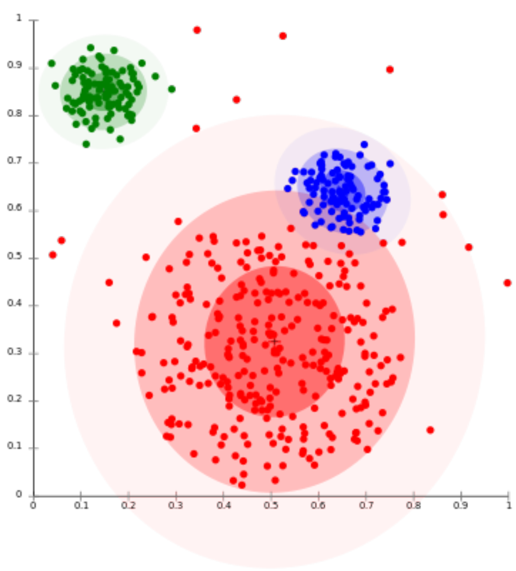
- HIERARCHICAL: finds successive clusters using previously established clusters.
- PARTITIONAL: divide data objects in non overlapping subsets.

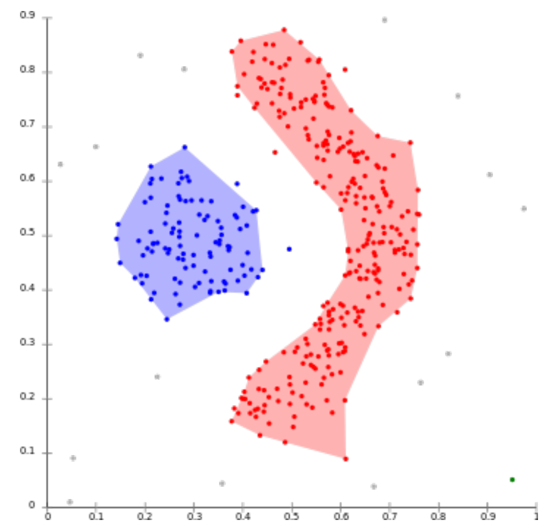




Types of Clustering

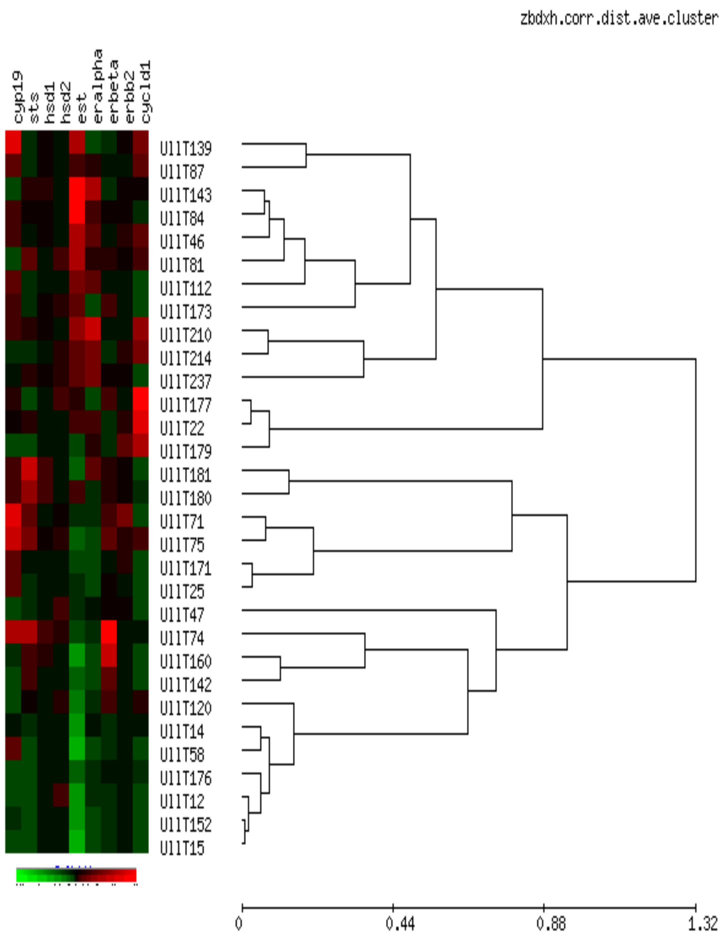
- MODEL BASED: assumes that the data were generated by a model and tries to recover the original model from the data (e.g., EM).
- DENSITY BASES: clusters are defined as areas of higher density. Objects in these sparse areas are usually considered to be noise and border points (e.g., DBSCAN).





Hierarchical Clustering

- Find subsequent clusters using previously established ones.
- Cannot test all possible trees.
- Agglomerative (bottom-up): we start with each element in a separate cluster and merge them accordingly to a given property.
- Divisive (top-down): start with all the points in the same clusters and then divide them.



Dendogram

- Hierarchical clusters are visualized as a dendogram.

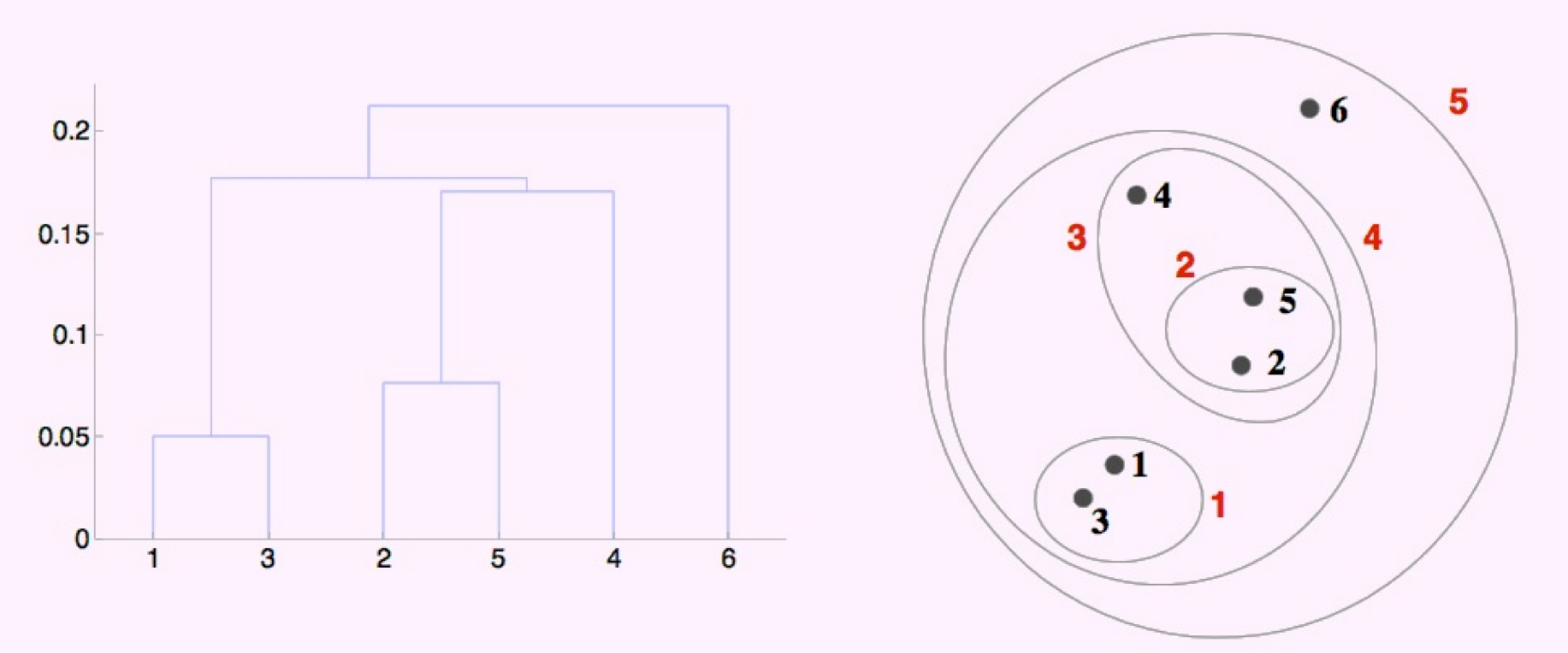


Image credits: Matteo Pardo

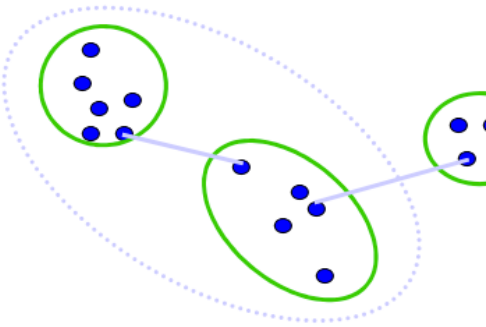
Hierarchical Clustering: pro and cons

- Pro
  - no need to specify the number of clusters;
  - intuitive.
- Cons
  - scalability [$O(n^2)$, n=# of samples];
  - local optima;
  - subjective.

Distance between clusters

- Single link
  - smallest distance between an element in one cluster and an element in the other

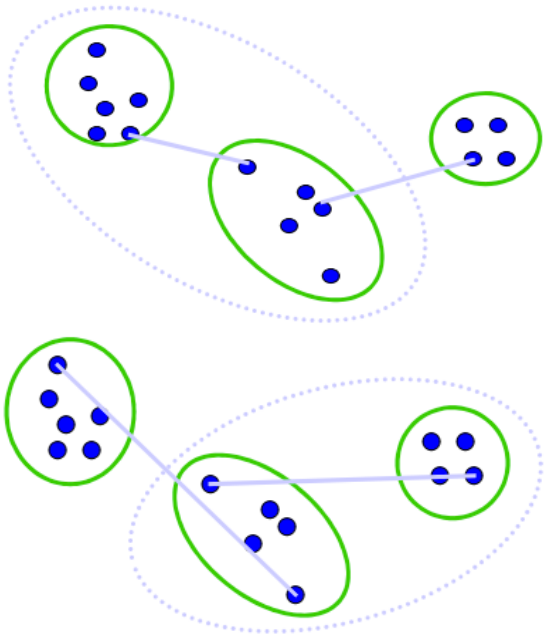
    $dis(K_i, K_j) = min(t_{ip}, t_{jq})$



Ref: Data Mining: Concepts and Techniques, J. Han, M. Kamber
Image credit: Henry Lin

Distance between clusters

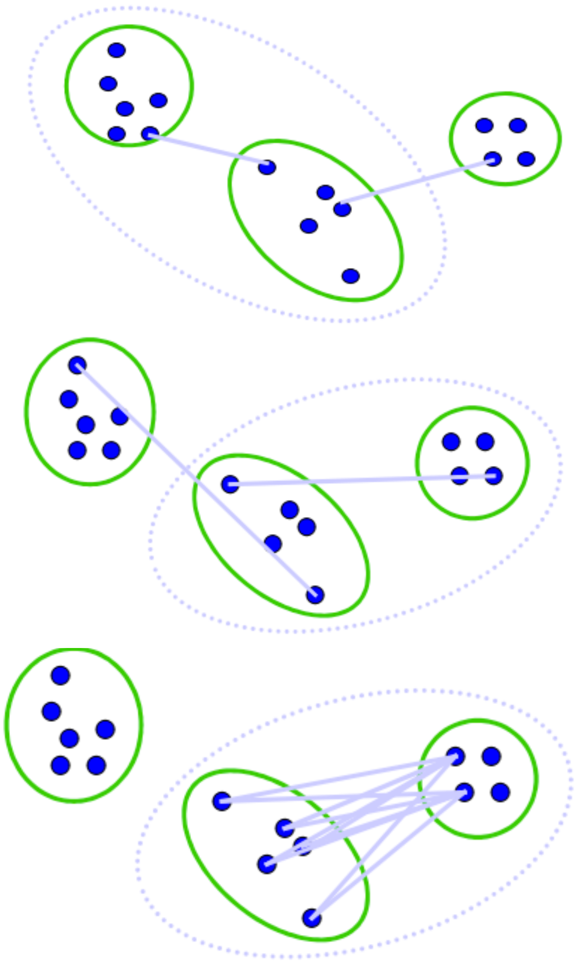
- Single link
  - smallest distance between an element in one cluster and an element in the other
    $dis(K_i, K_j) = min(t_{ip}, t_{jq})$
- Complete link
  - largest distance between an element in one cluster and an element in the other
    $dis(K_i, K_j) = max(t_{ip}, t_{jq})$



Ref: Data Mining: Concepts and Techniques, J. Han, M. Kamber
Image credit: Henry Lin

Distance between clusters

- Single link
  - smallest distance between an element in one cluster and an element in the other
    $dis(K_i, K_j) = min(t_{ip}, t_{jq})$
- Complete link
  - largest distance between an element in one cluster and an element in the other
    $dis(K_i, K_j) = max(t_{ip}, t_{jq})$
- Average
  - average distance between an element in one cluster and an element in the other
  - i.e., $dis(K_i, K_j) = avg(t_{ip}, t_{jq})$
- Centroid, Medoid



Ref: Data Mining: Concepts and Techniques, J. Han, M. Kamber
Image credit: Henry Lin

Similarity Measures

- Determine the similarity between two clusters and the shape of the clusters.
- Based on distance metrics.
- Distance metric: defines a distance between elements of a set:
  - non negativity: $dist(x,y) \geq 0$
  - symmetry: $dist(x,y) = dist(y,x)$
  - self-similarity: $dist(x,y)=0 \Leftrightarrow x=y$
  - triangular inequality: $dist(x,z) \leq dist(x,y) + dist(y,z)$

Common Distances

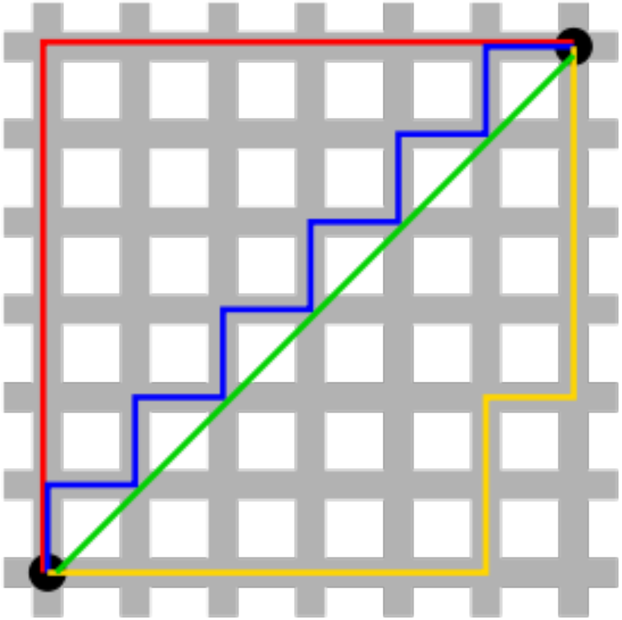
- Euclidian Distances
  - commonly used;
  - sphere shaped clusters;
  - Squared Euclidean Distance is not a metric as it does not satisfy the triangle inequality, however it is frequently used in optimization problems.

Names	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\|a-b\|_1 = \sum_i \lvert a_i - b_i \rvert$
maximum distance	$\|a-b\|_\infty = \max_i \lvert a_i - b_i \rvert$
Mahalanobis distance	$\sqrt{(a-b)^\top S^{-1}(a-b)}$ where S is the covariance matrix
cosine similarity	$\frac{a \cdot b}{\|a\|\|b\|}$

Common Distances

- Taxicab distance (Manhattan distance)
  - distance between two points is the sum of the absolute differences of their Cartesian coordinates;
  - “diamond” shaped clusters;

Names	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
maximum distance	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$
Mahalanobis distance	$\sqrt{(a-b)^\top S^{-1}(a-b)}$ where S is the covariance matrix
cosine similarity	$\frac{a \cdot b}{\|a\|\|b\|}$



Common Distances

- Mahalanobis Distance
  - measure of the distance between a point P and a distribution D.
- Cosine Similiarity
  - commonly used in information retrieval;
  - e.g., in text mining gives a useful measure of how similar two documents are likely to be in terms of their subject matter.

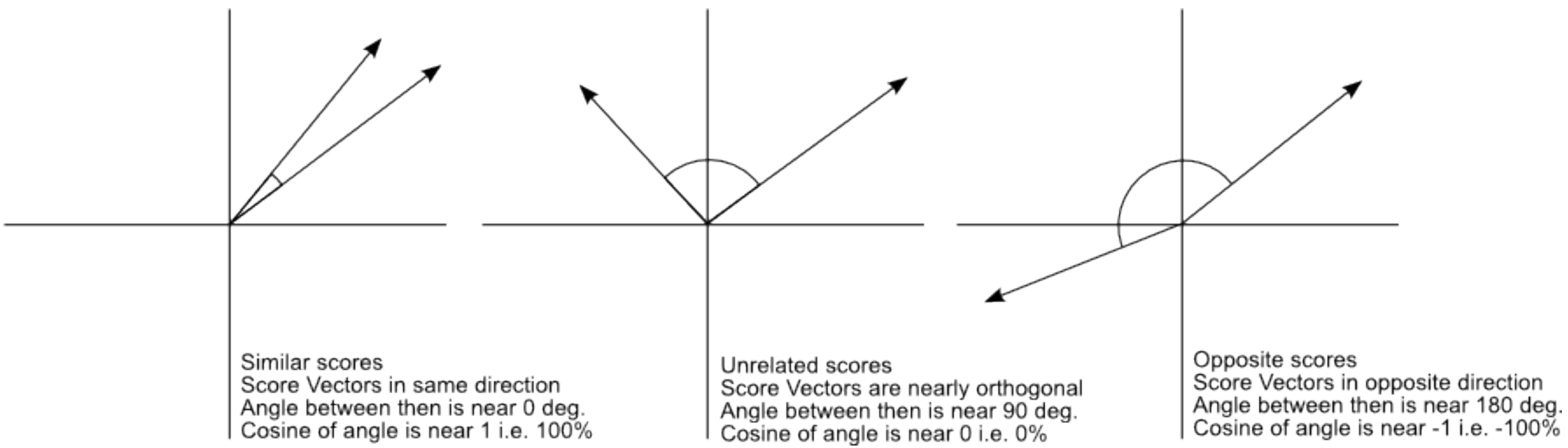
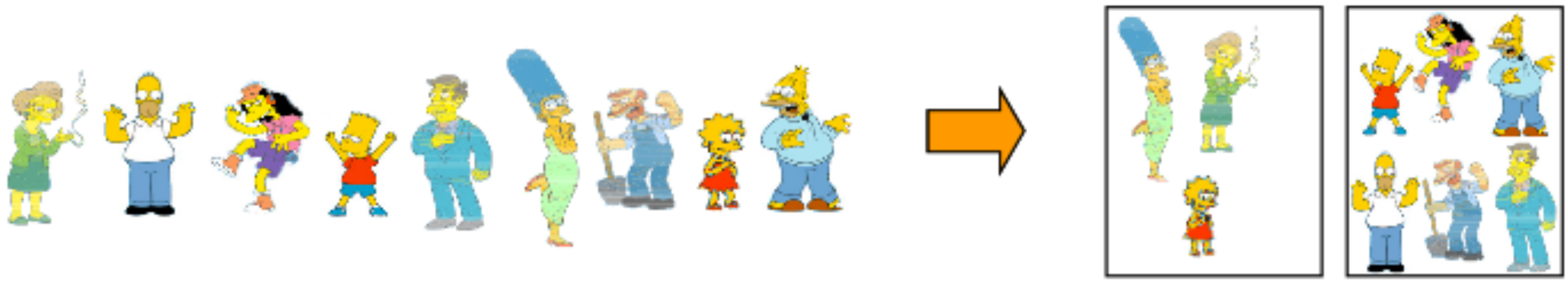


Image credits: pyevolve

Partitional Clustering

- Flat clustering.
- Need to specify or compute k.
- Each instance is placed in one of the clusters.
- Hard and Soft clustering.

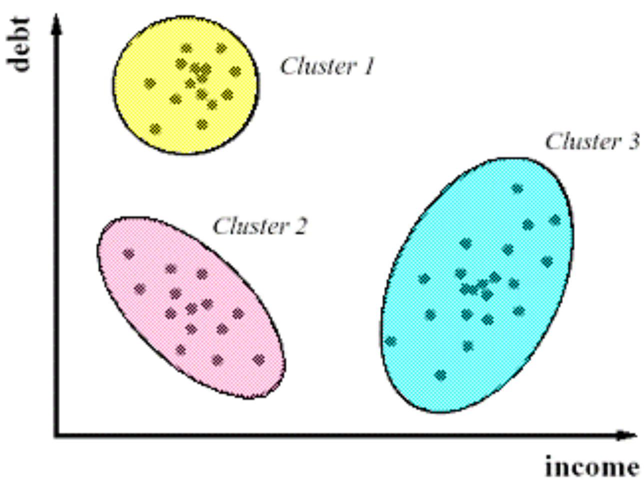


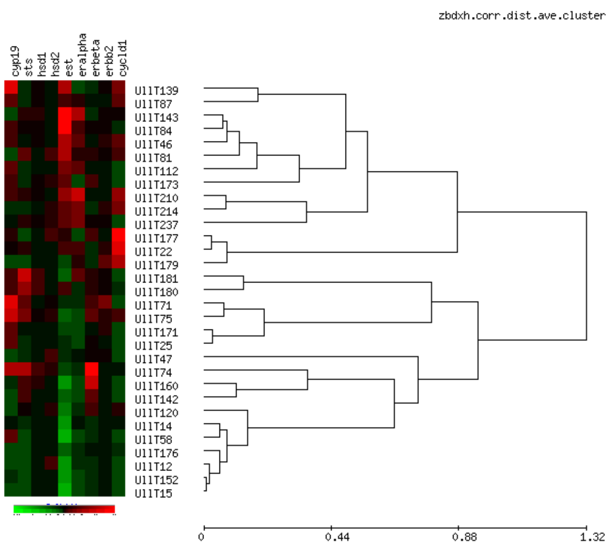
Hard and Soft Clustering

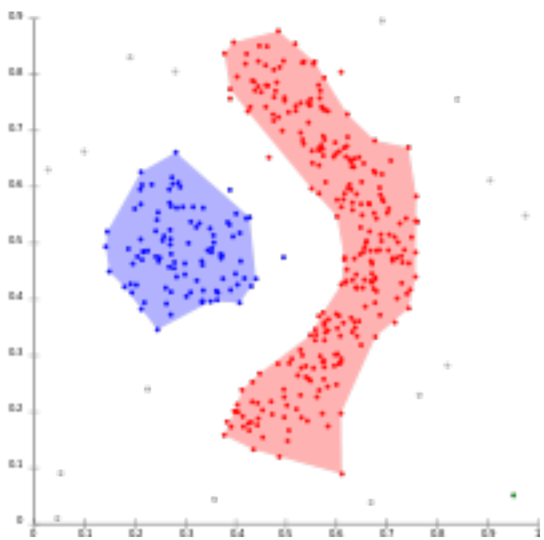
- Hard clustering
  - each sample is member of one cluster exactly.
  - e.g., k-means
- Soft clustering
  - each sample has a degree of membership for each cluster.
  - e.g., Expectation-Maximization (EM)
- Exhaustive vs Non-exhaustive clustering

Summary

- Clustering
  - model
  - metric
  - objective function
  - hard or soft







[redacted]