# Highly Dimensional Data and Manifold Learning

# Dimensionality Reduction

- Clusters may exist in a subspace of the dimensions
- Two data points may belong in a cluster in one set of dimensions, but not in another set of dimensions
  - Known as “local feature relevance”
- Chop out correlated or unrelated dimensions
  - Potentially lose clusters

# Axis Parallel: Types

- Projected Clustering
  - Each data point is either in exactly one cluster or is considered noise
- Soft Projected Clustering
  - Find the best k-clusters from the data
- Subspace Clustering
  - Find all subspaces that have a cluster

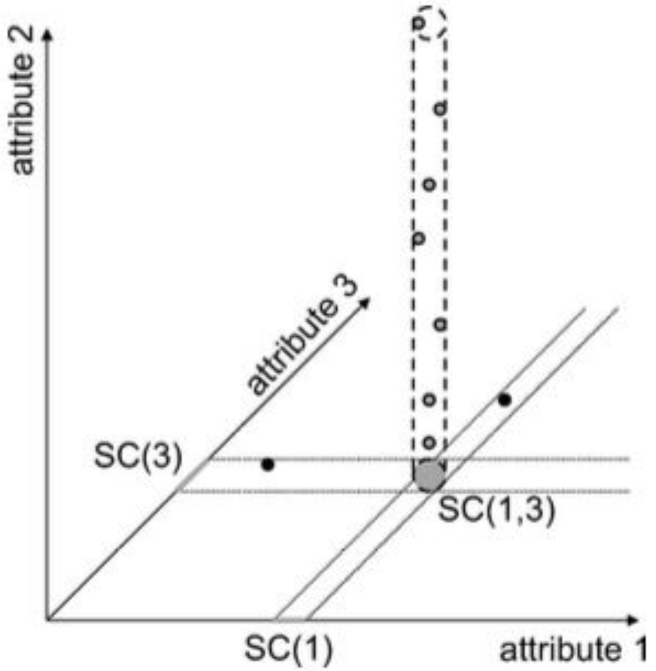
# Axis Parallel: Subspace Clustering

Top Down:

1. Select potential cluster members
    a. Locality assumption or random sampling
2. Sample the variance in each dimension
3. Select the subspace in which they meet the cluster requirement

Bottom Up:

1. Select a dimension
2. Determine if it has a cluster
3. If so, combine it with a new dimension



# Data Matrix Clustering

Data is represented in a matrix A(data, dimensions)

- Constant Biclusters: recursively split the matrix in two, maximizing the reduction in variance (inefficient)
- Biclusters with Coherent Evolutions: find the largest subset of the matrix, such that there exists a permutation of the columns where the values in each row is strictly increasing
    - The quality of the subset is measured by the number of rows fitting the inequality

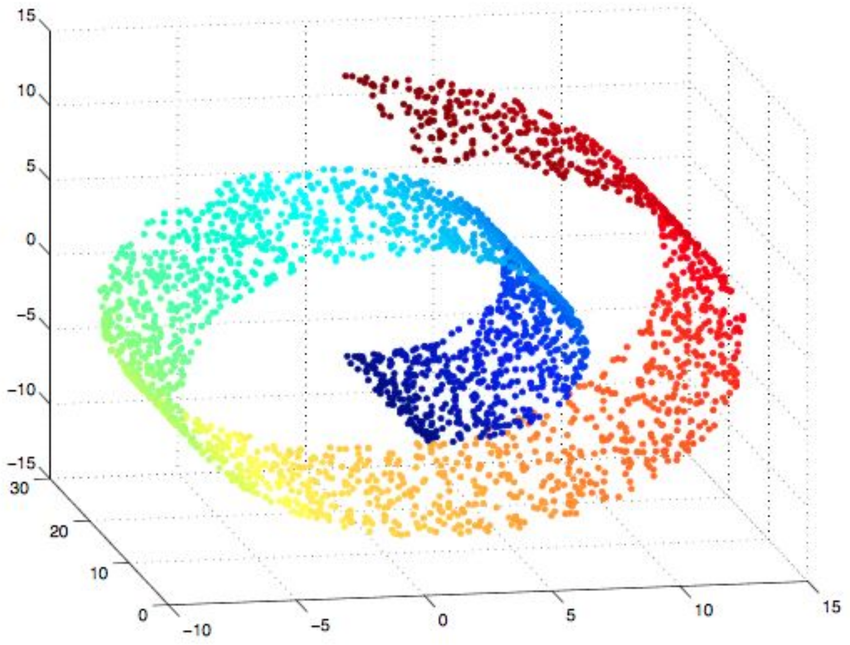
# Manifold Learning

Manifold: an object which locally resembles euclidean space at each point

Manifold Learning: discovering which dimensions matter in a high-dimension data set

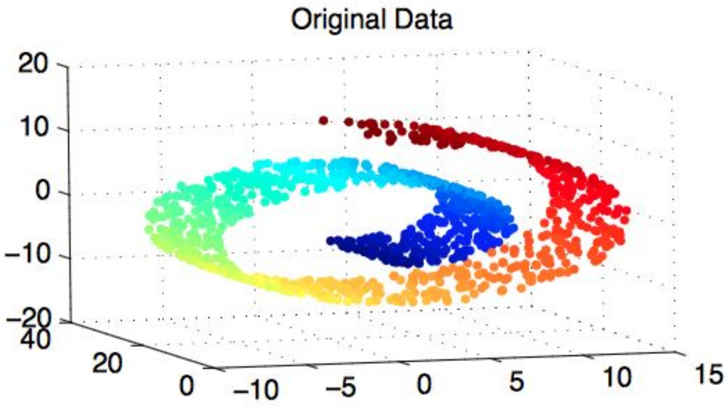
# Principal Components Analysis (PCA)

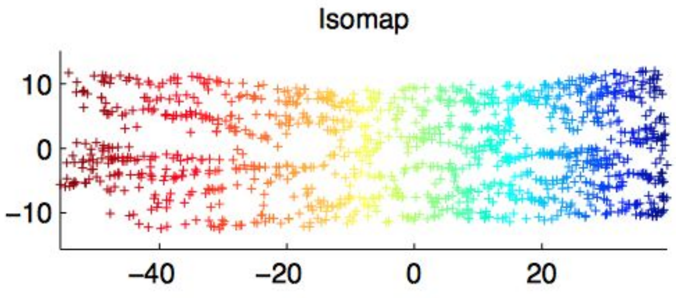
- Finds vectors where data has maximum variance
  - Allows for vectors not parallel to axes
- Vectors are weighted based on amount of variance (importance)
- Works well for linear subspaces



# IsoMap

1. Estimate the distance along the manifold between points
    - Uses direct distance for local points
    - Uses shortest path on a nearest neighbor graph (such as Dijkstra's) for further points
2. Use multi-dimensional scaling to find matching points in a low-dimension Euclidean space



# Visualization Tools: RadViz

- Based on a unit circle, with “springs” between dimensions and the data point
- Each to see relations between each dimension
- Loses specific data about each dimension
- Clusters become more evident

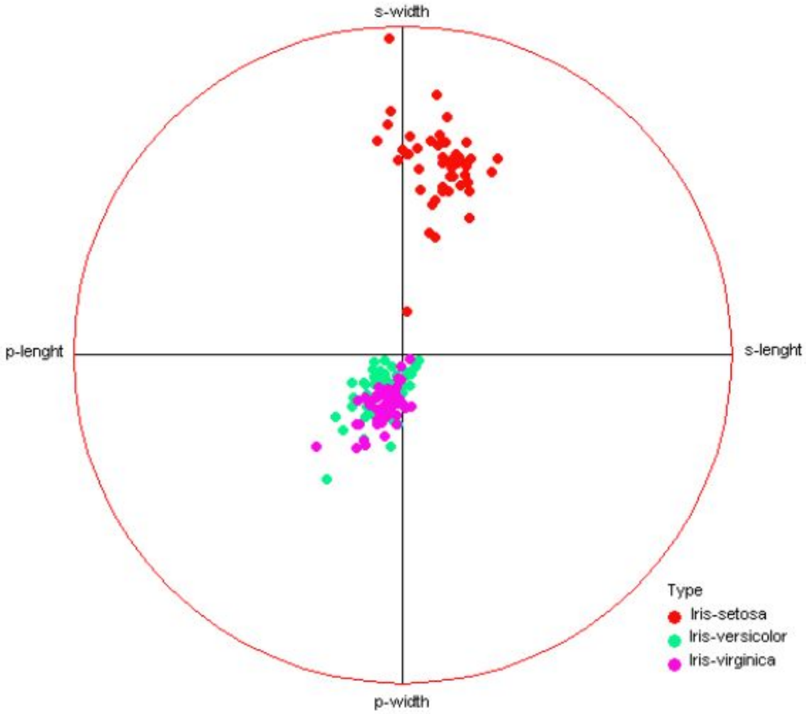



**Figure 3.15:** RadViz visualization of the Iris data set

# Visualization Tools: Parallel Axes

- Excellent for viewing whole data set
- Obvious relationships between consecutive axes
  - Can be difficult to see relationships between further axes
- Can be difficult to follow a single data point
- Can also be displayed as a circular polar chart

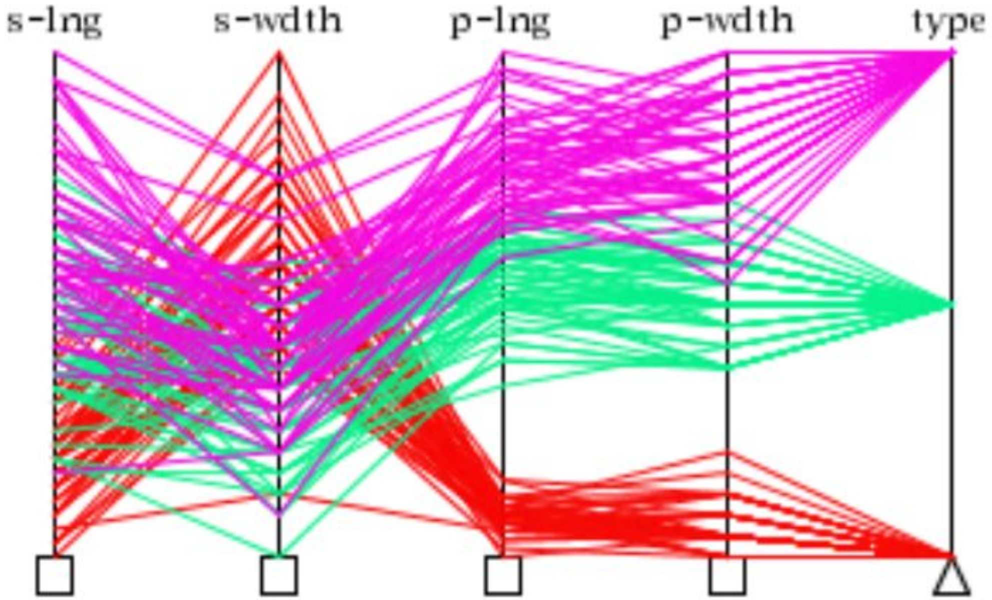
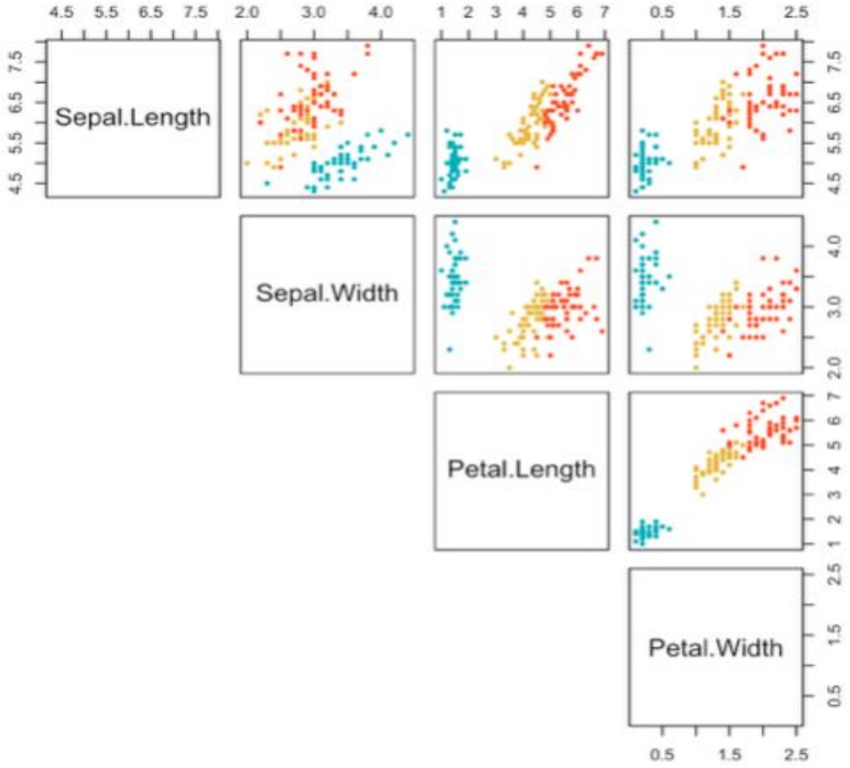



**Figure 3.10:** Parallel coordinate display of the Iris data set

Image from: http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs

# Visualization Tools: Scatter Plot Matrices

- Maintains exact values of each dimension
- Easy to examine relationships between dimensions
- Easy to notice clusters in different dimensions
- Difficult to follow single data point



# References

[1] Kriegel, H.-P., Kroger, P., and Zimek, A. 2009. Clustering high-dimensional data: A survey on sub-space clustering, pattern-based clustering, and correlation clustering. ACM Trans. Knowl. Discov. Data. 3, 1, Article 1 (March 2009), 58 pages. DOI = 10.1145/1497577.1497578 http://doi.acm.org/10.1145/1497577.1497578

[2] Cayton, L. 2005. Algorithms for manifold learning. http://www.vis.lbl.gov/~romano/mlgroup/papers/manifold-learning.pdf

[3] Grinstein, G., Trutschl, M., Cvek, U. High-Dimensional Visualizations. Institute for Visualization and Perception Research. https://pdfs.semanticscholar.org/43f7/66c06e2a7770d9f37dcd9cfff5bd5dcfc22f.pdf