“No Free Lunch” Theorem

Masha Kinley

- Clustering
    - Applications
    - Advantages and disadvantages
    - Categories
- Algorithms
    - K-Means
    - Hierarchical (hclust)
    - Density based (DBSCAN)
- Experiments and results

Clustering

- Division of data into groups of similar objects
- Each cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups
- Reveal hidden patterns

Clustering as a data mining tool:

  - Biology
  - Medicine
  - Security
  - Business intelligence
  - Web search

- Powerful tool but requires planning and preparation

Clustering methods

- Partitioning
  - “One object – one group”. Most are distance based. Spherical shape.
- Hierarchical
  - Bottom-up or Top-down. Cannot be undone.
- Density-based
  - Number of objects in the neighborhood. Arbitrary shapes.
- Grid-based
  - Fast processing time. Grid size matters.

Algorithms

- K-Means
  - Centroid of each cluster represents that cluster
    - Centroid – mean value of the objects in the cluster
    - Centroid is randomly selected
    - Euclidean distance is then measured between each other object and the cluster mean
    - Iterations improve within-cluster variations and new means are assigned
    - Iterations continue until the clusters are stable between iterations
  - Fast computing speed
  - Does not deal with non-convex shapes
  - Will assign outliers to a cluster
  - Number of clusters as an input parameter

Algorithms

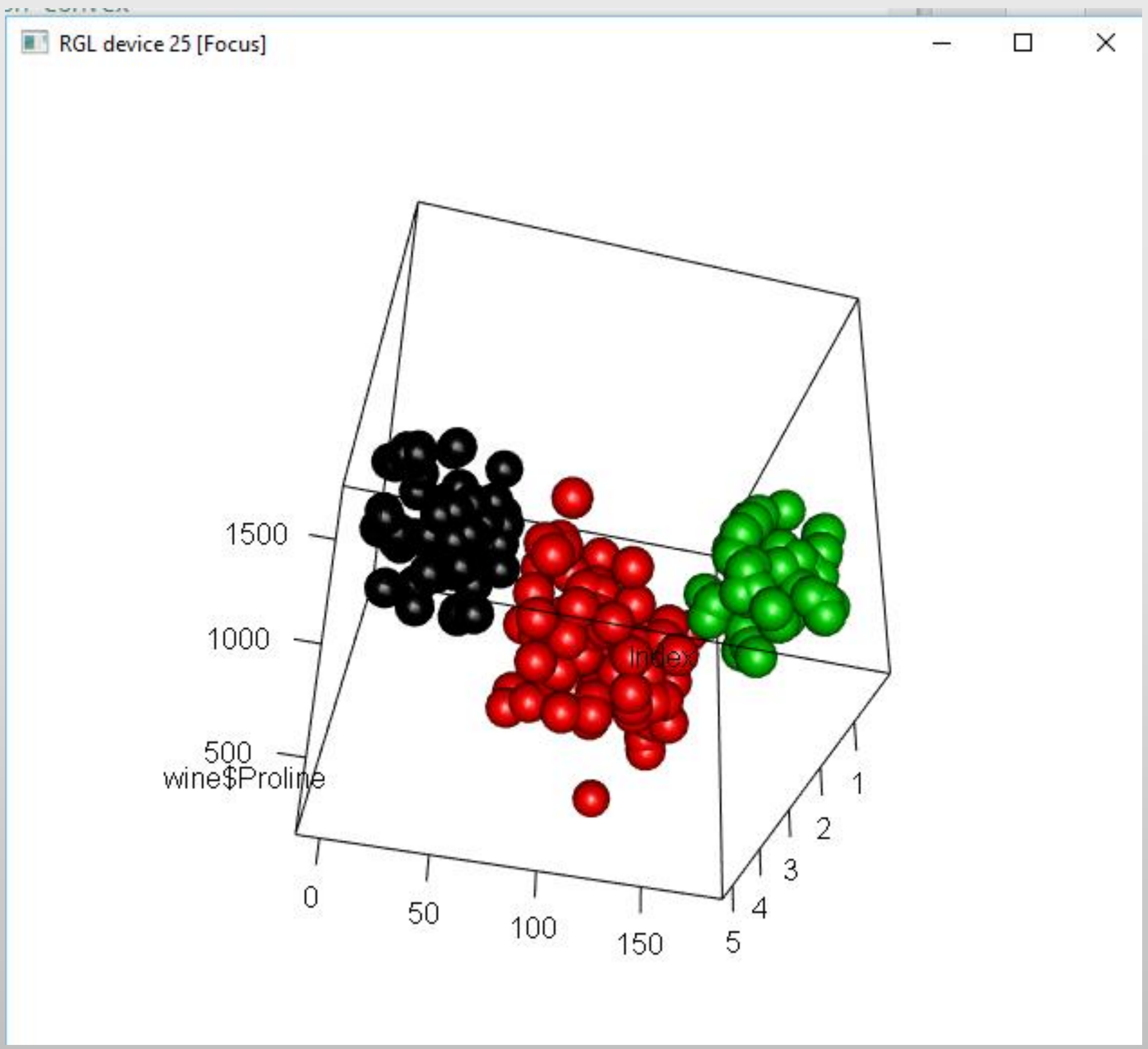
- Hierarchical
  - Forms a “tree” of clusters – a dendrogram
  - Useful for data summarization or visualization
  - Distance between clusters of objects
  - Many types
    - hclust
      - Bottom-up – each point is its own cluster
      - Closest two clusters are combined into one
      - Repeats until all points are one cluster
  - Can be too sensitive to outliers
  - Difficult to interpret results for large datasets

Algorithms

- DBSCAN
  - Based on connected regions of high density
  - Mass/volume
    - Point *p* and its neighborhood of radius ε, the
    - *mass* of the neighborhood number of data points contained within such neighborhood
    - *volume* of the neighborhood is volume of the resulting shape of the neighborhood thus defining the density at the point *p* of the given neighborhood.
    - Core points, border points and outliers
  - Time and computing power
  - Poor clustering quality when data density is uniform
  - Input parameters (radius and min points) are hard to determine

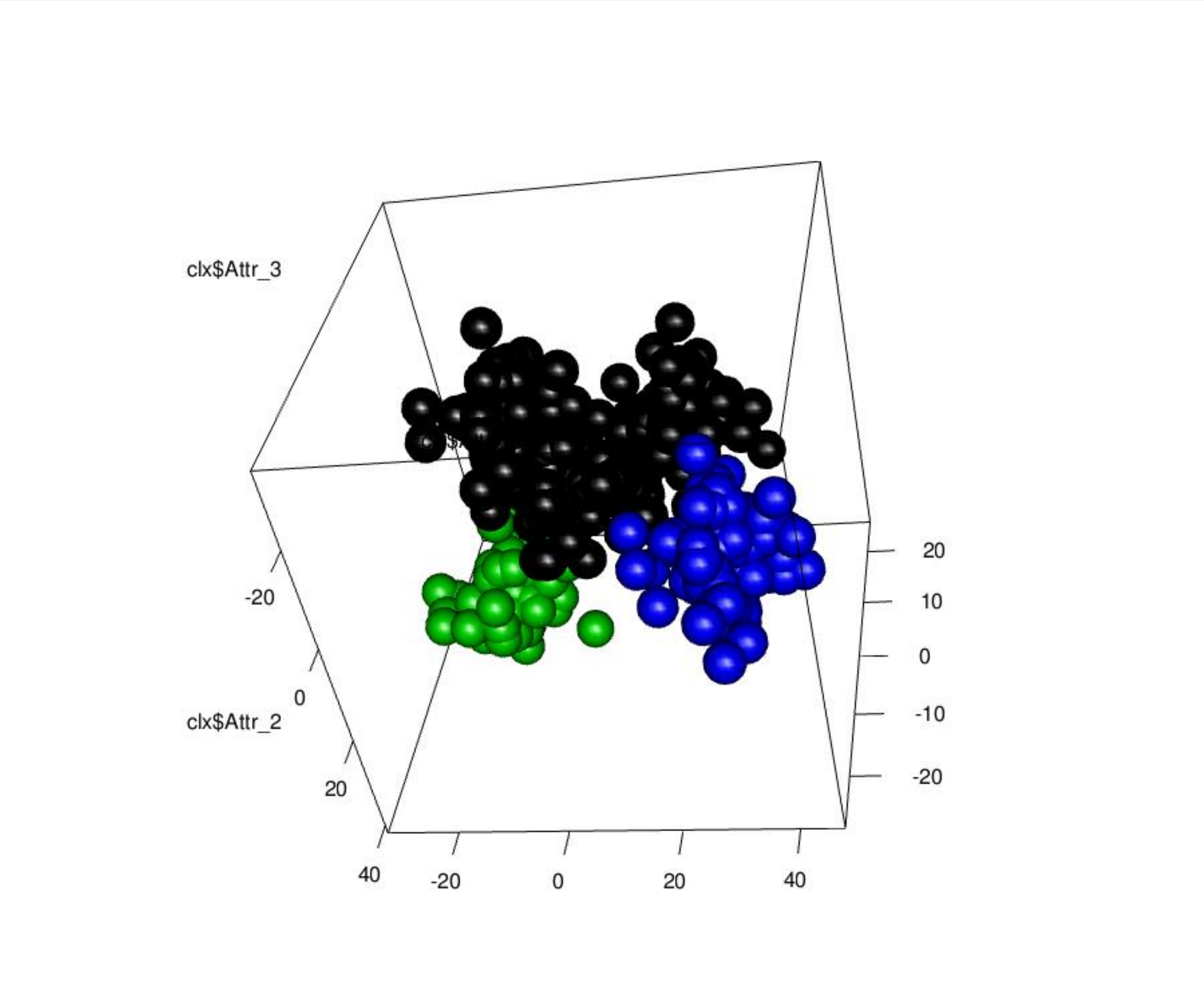
Experiment

- Two datasets
  - Wine
    - 3 distinct spherical clusters
    - 178 instances, 13 attributes
    - No missing values
    - very little noise



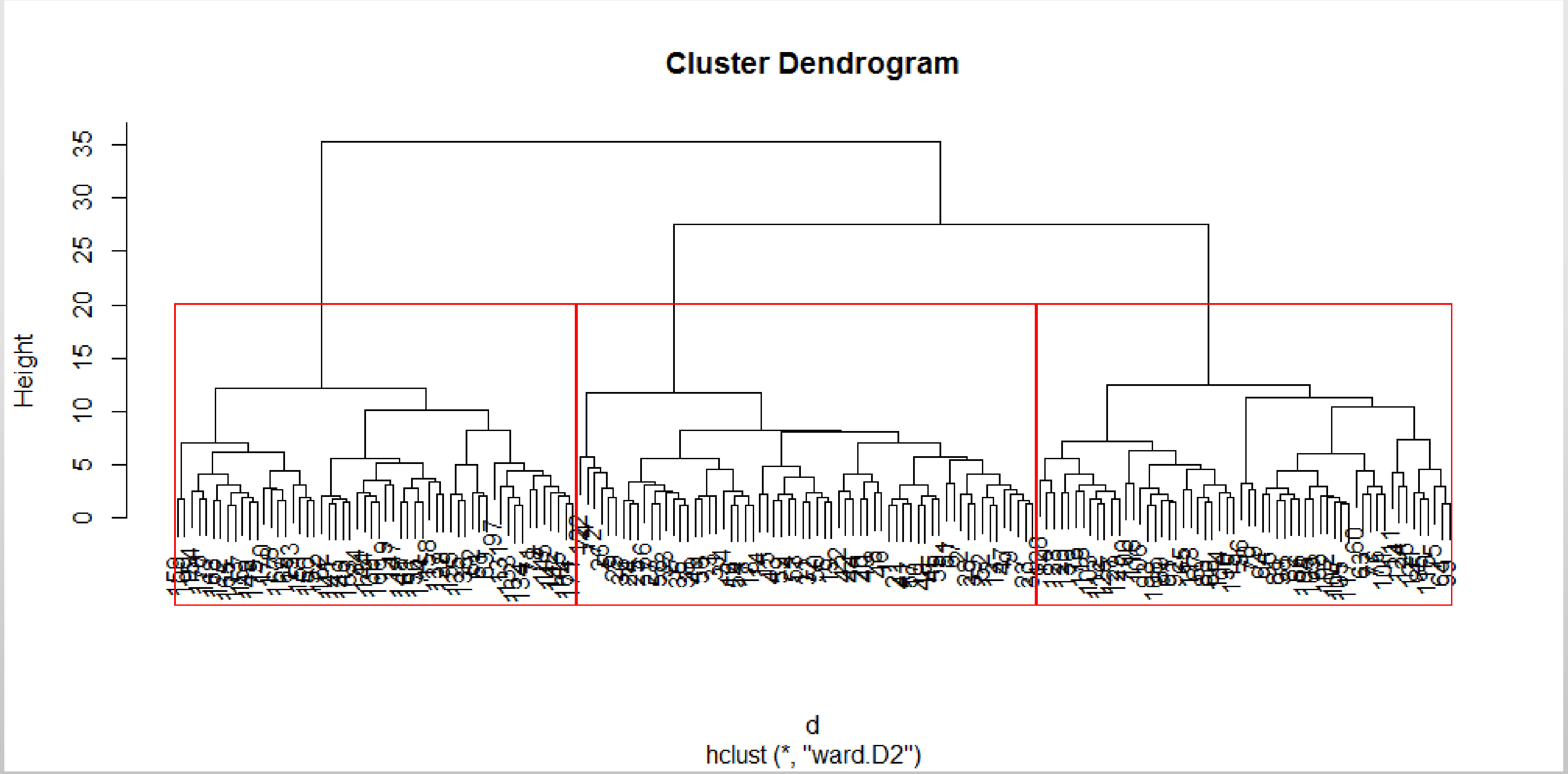
Experiment

- Clx Dataset
  - Created by Dr. Aleshunas
  - 3 non-convex clusters
  - 827 instances
  - 3 attributes



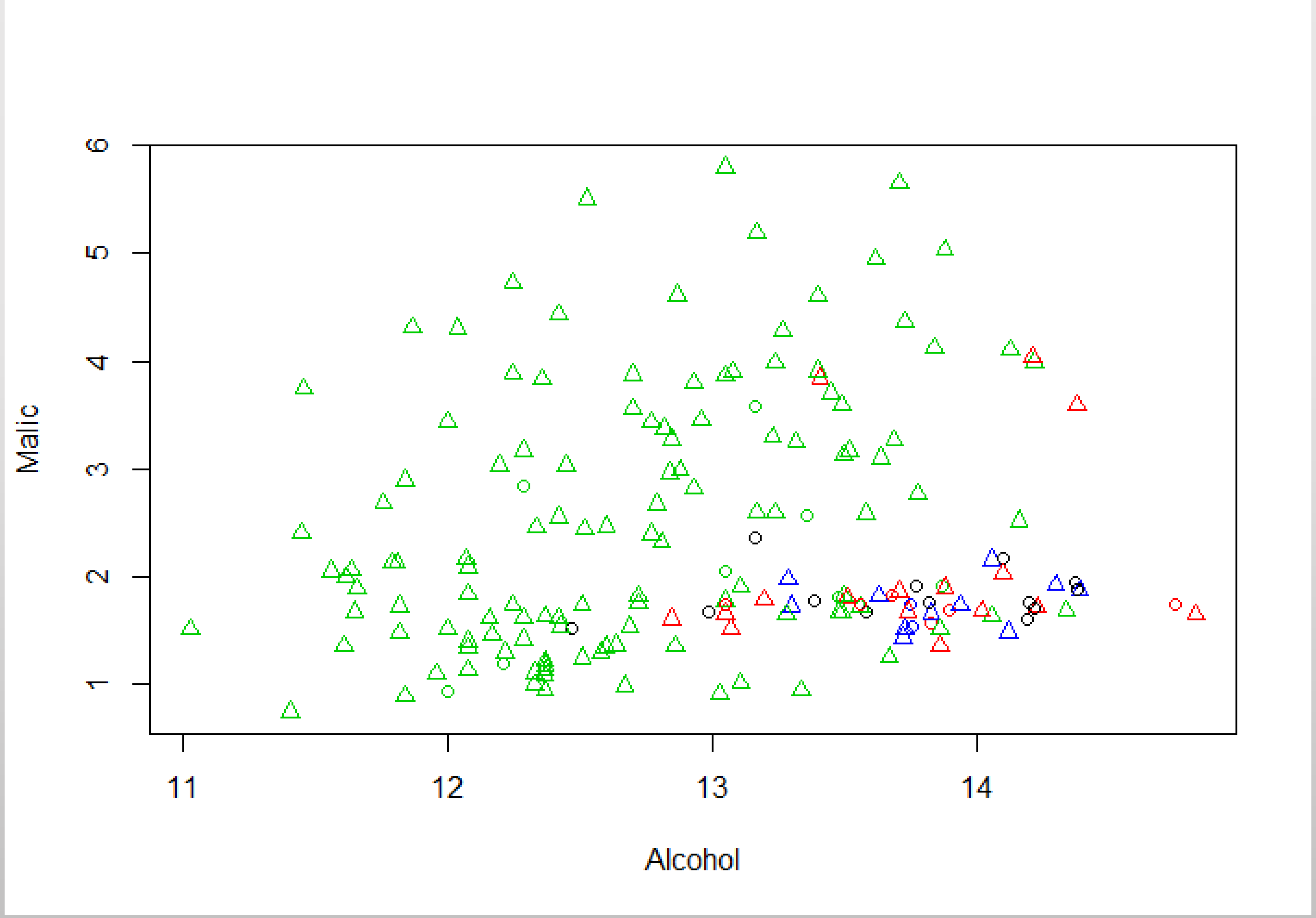
Results – Wine

- Hclust



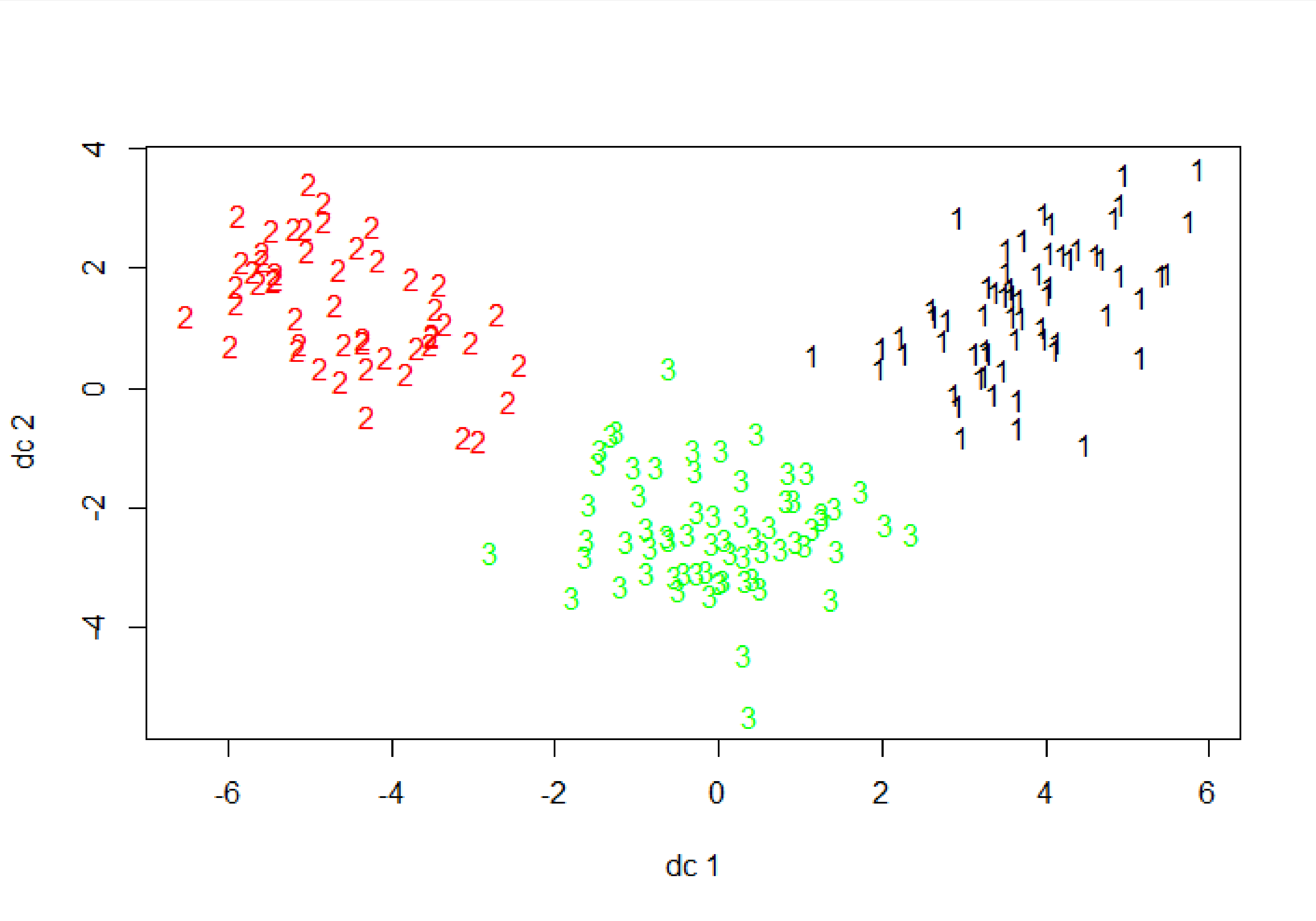
Results – Wine

- DBSCAN



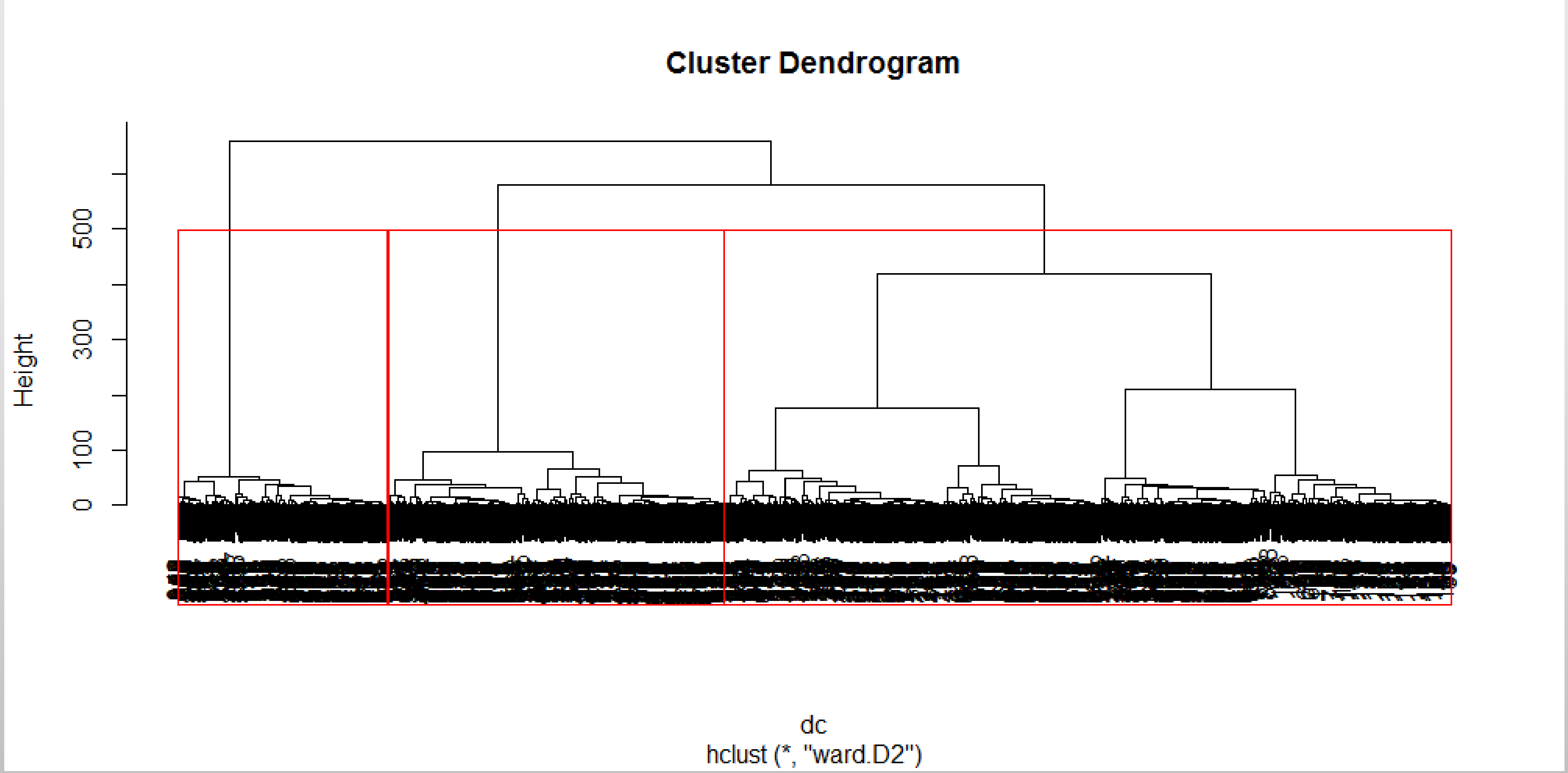
Results – Wine

- K-Means



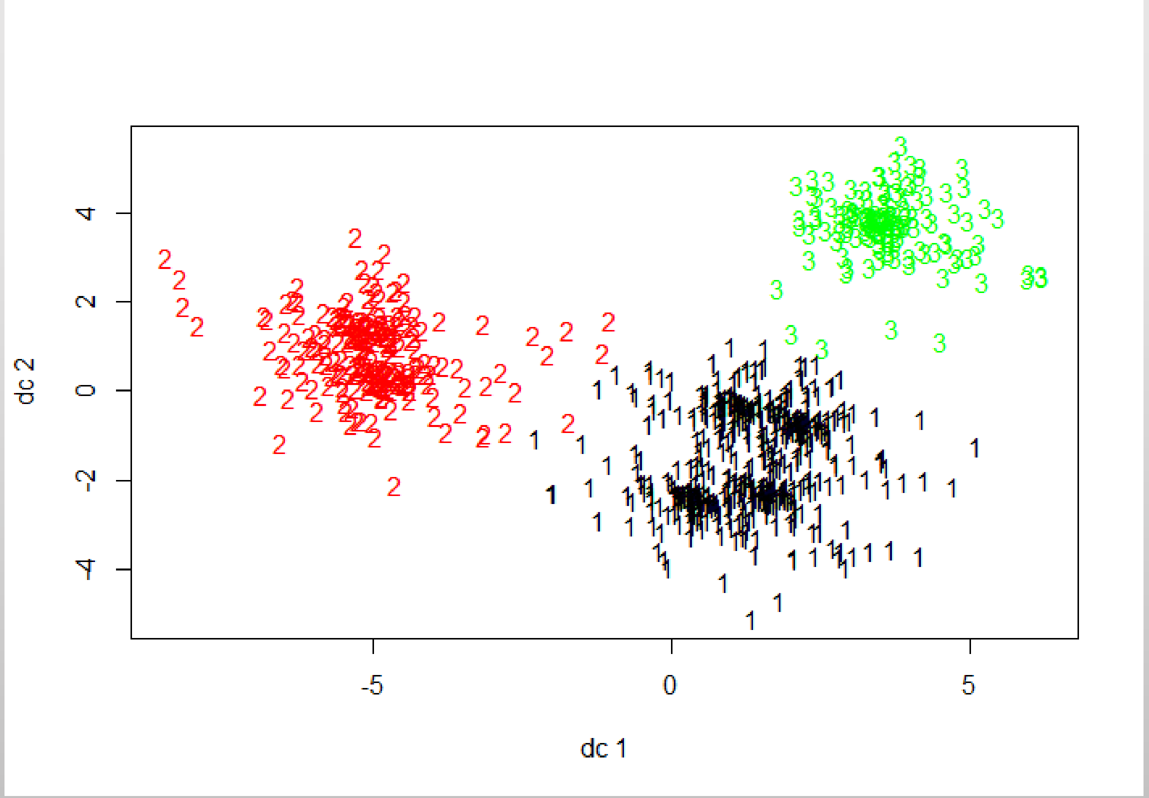
Results – Clx

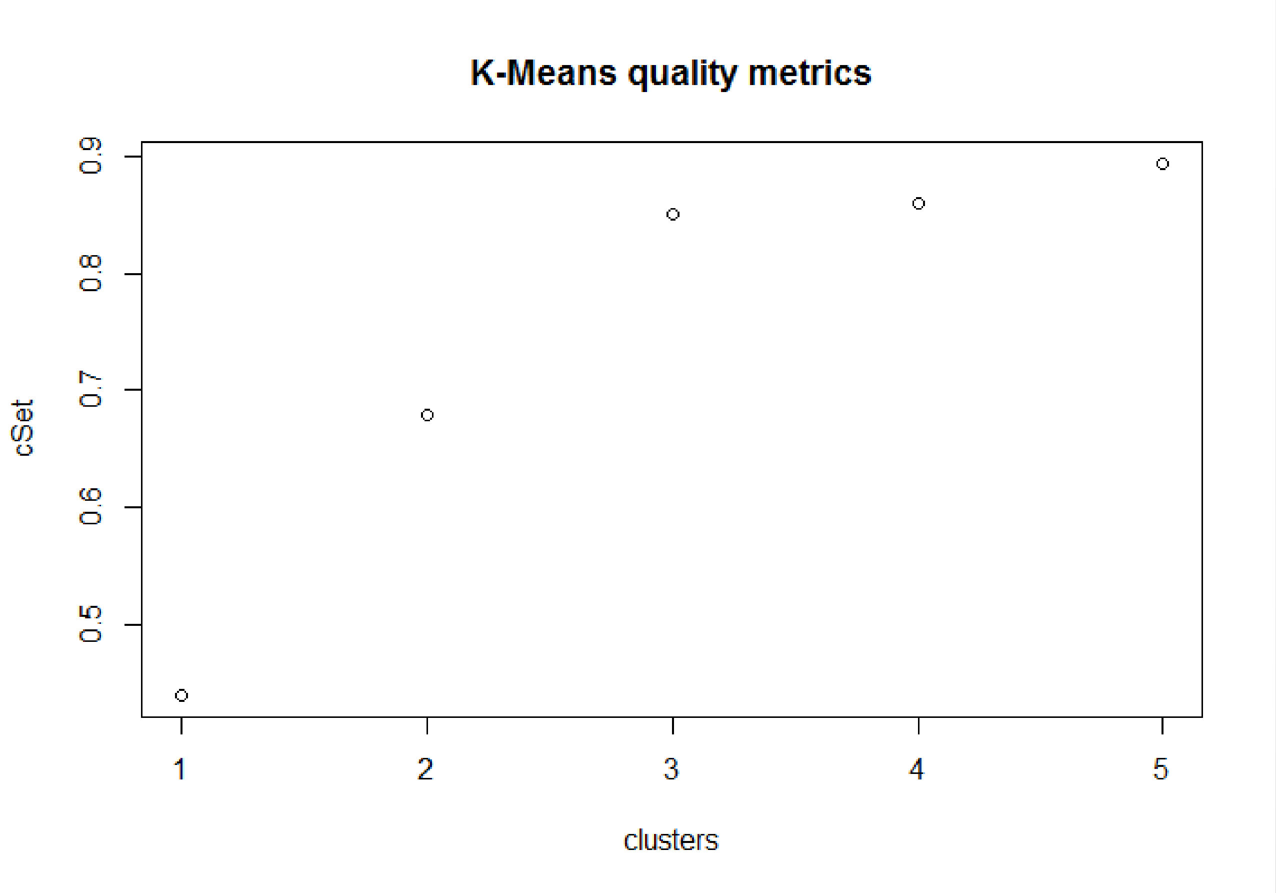
- Hclust



Results – Clx

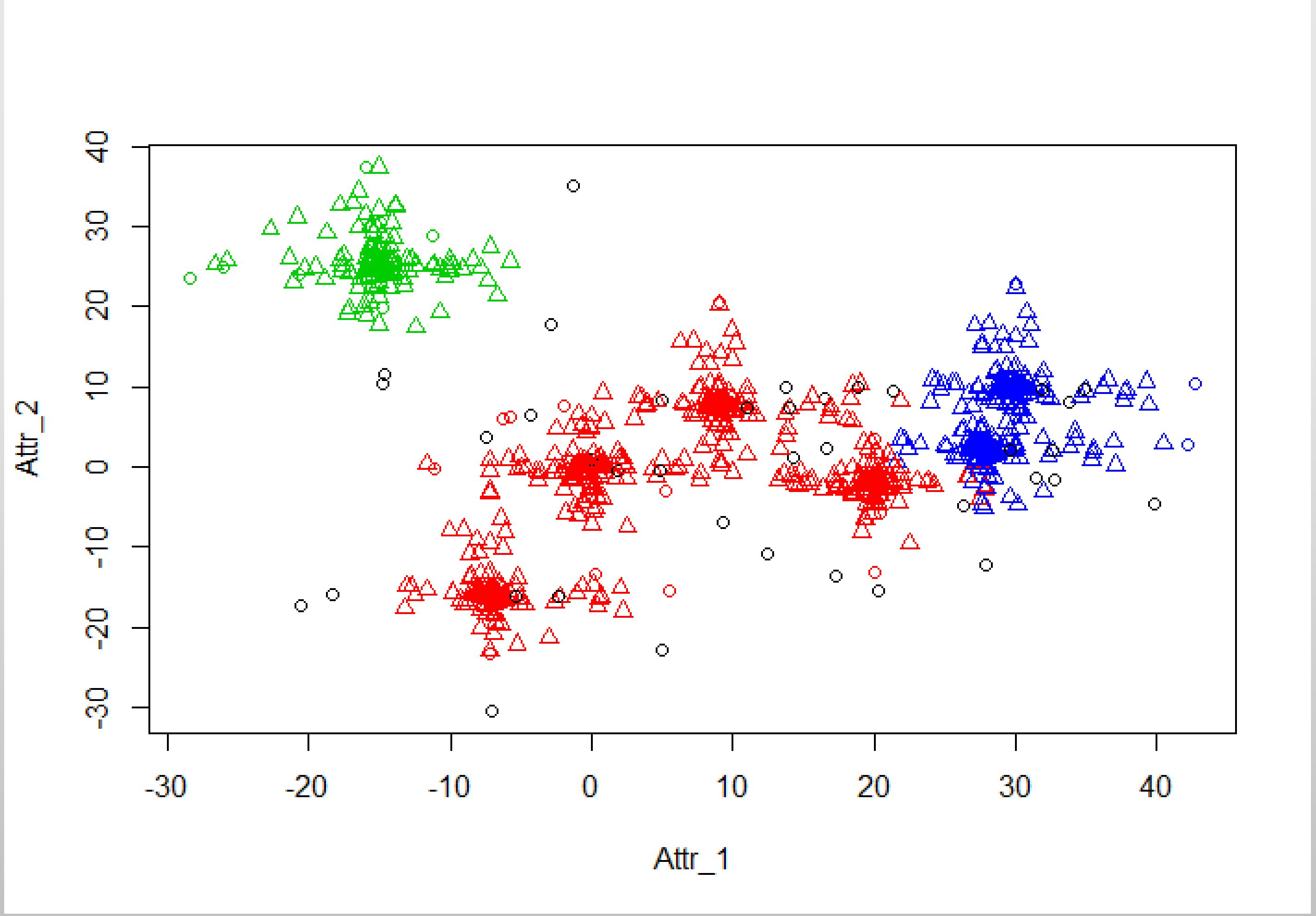
- K-Means





Results – Clx

- DBSCAN



Conclusion

- No free lunch
  - multiple methods should be explored in each case
  - nature of the dataset must be considered
- Questions?

Sources

- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques .* Waltham: Morgan Kaufman .
- Kodali, T. (2016, January 22). *Hierarchical Clustering in R.* Retrieved from R-Bloggers : https://www.r-bloggers.com/hierarchical-clustering-in-r-2/
- Galili, T. (2015, August). *K-Means Clustering .* Retrieved from R-Statistics : https://www.r-statistics.com/2013/08/k-means-clustering-from-r-in-action/