# [[ Data Mining ]]

## 01-Clustering
## Connectivity-Based Approach

# Cluster Analysis

Clustering is a technique used for combining observations into groups or clusters such that

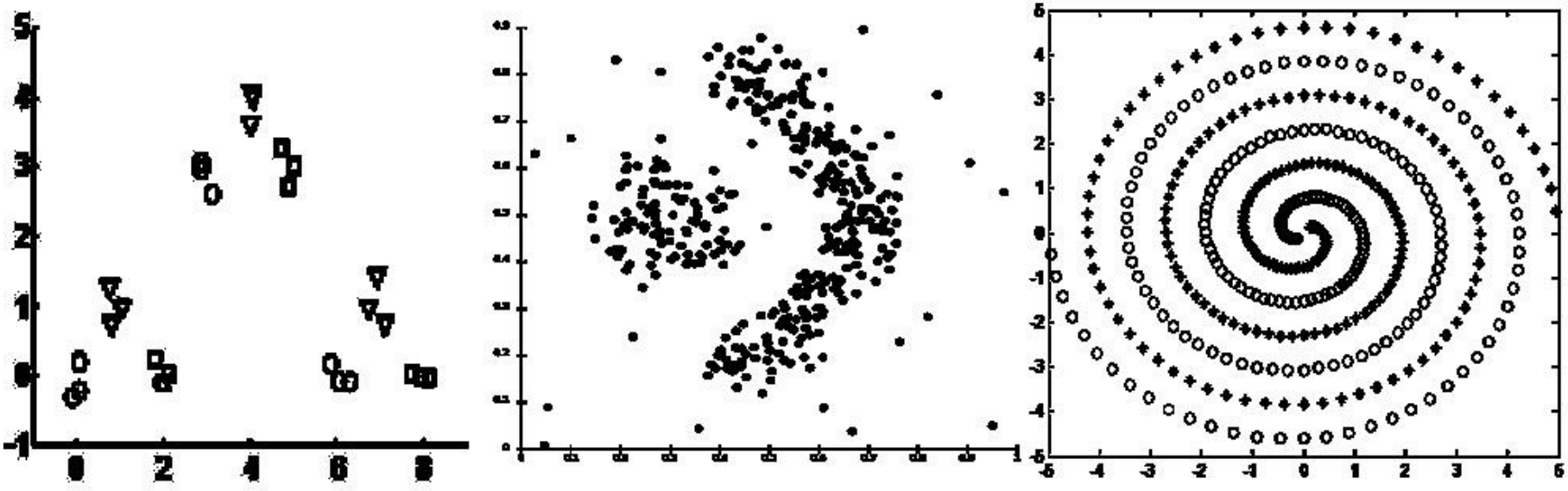
- Each group or cluster is homogeneous or compact with respect to certain characteristics.
- Each group should be different from other groups with respect to the same characteristics.

# Two-Dimensional Data Example

| ID | x | y |
|---|---|---|
| 1 | x1 | y1 |
| 2 | x2 | y2 |
| 3 | x3 | y3 |
| … | … | … |
| n | xn | yn |

- How similar is “similar”?
- How to define “similarity”?

2. 這個就比較難分，很難說彎彎的是同一群因為y差太多



3.這也很難說這可以分成兩群

# Non-Structured Data Examples

- Which two sentences are more similar?

To infinity and beyond!

要想一下feature 到底是什麼

Mirror, mirror on the wall,
who is the fairest of them all?

text mining 可以去研究一下

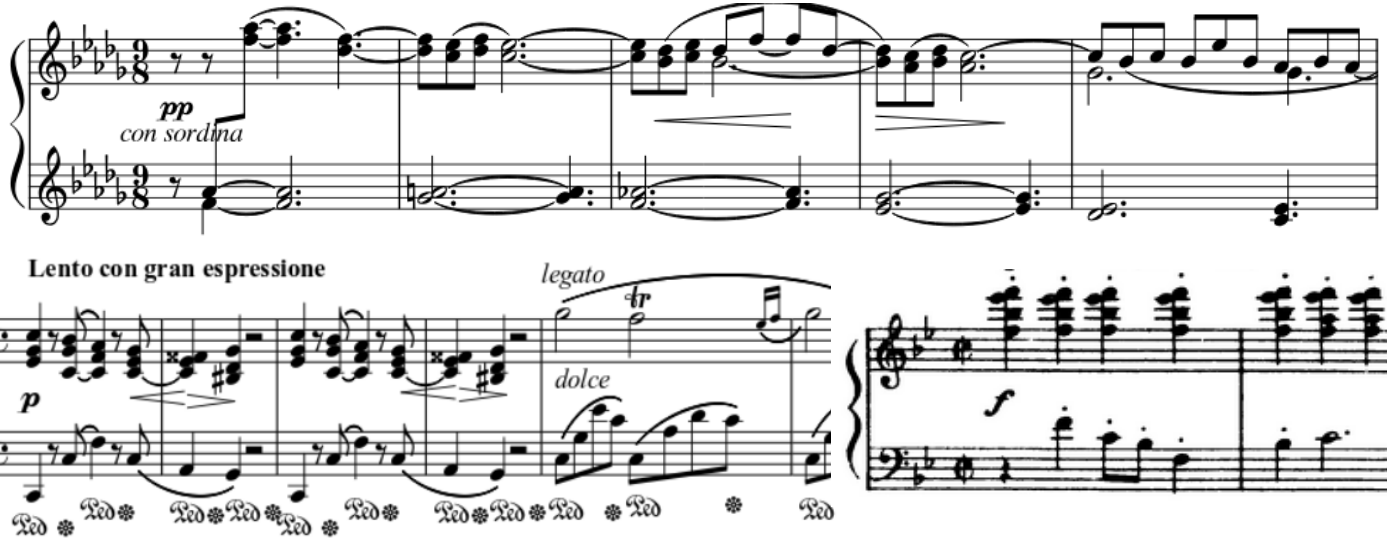
Let it go, let it go,
I can't hold it anymore.

- Which two photos are more similar (in age)?



texture analysis
LBP

- Which two pieces are more similar (in genre)?



## Clustering Algorithms

- Connectivity-based: Hierarchical Clustering
- Partition-based: K-means and K-medoids
- Density-based: DBSCAN

## Extensions to Large-Scale Data

- Hierarchical Clustering
- K-medoids

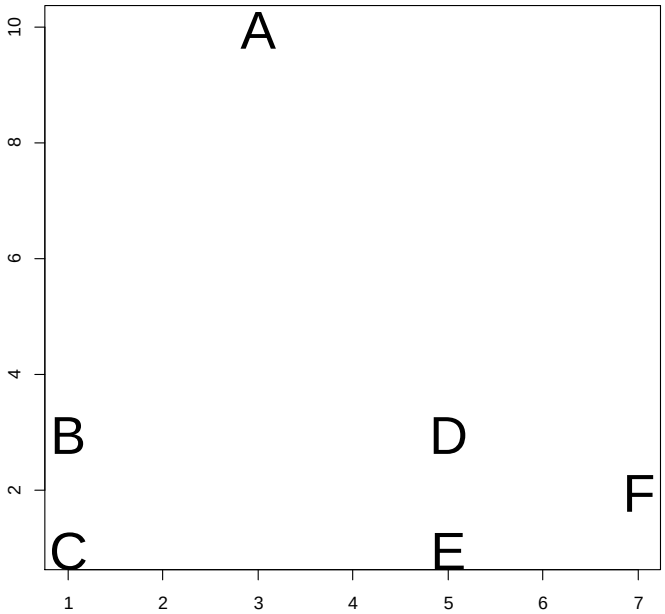
## Feature Extraction + Clustering

- Spectral Clustering
- For Categorical and Functional Data

# Hierarchical Clustering

這屬於connective

# Example




| ID | x | y |
|---|---|---|
| A | 3 | 10 |
| B | 1 | 3 |
| C | 1 | 1 |
| D | 5 | 3 |
| E | 5 | 1 |
| F | 7 | 2 |

- Agglomerative (bottom-up)
- Divisive (top-down)

# Agglomerative Hierarchical Clustering

Given a set of N items to be clustered.

1. Assign each item to its own cluster.
2. Compute distances (similarities) between clusters.
3. Find the closest (most similar) pair of clusters and merge them into a single cluster.
4. Repeat 2. and 3. until all items are merged into a single cluster.

# Algorithm Step 1 and 2

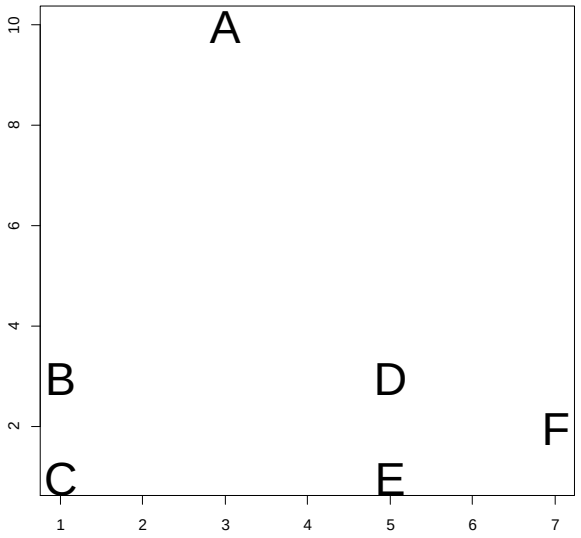
Given a set of N items to be clustered.

| ID | x | y |
|---|---|---|
| A | 3 | 10 |
| B | 1 | 3 |
| C | 1 | 1 |
| D | 5 | 3 |
| E | 5 | 1 |
| F | 7 | 2 |

1. Assign each item to its own cluster.

2. Compute the distances (dissimilarities) between clusters.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | | | | | | |
| B | | | | | | |
| C | | | | | | |
| D | | | | | | |
| E | | | | | | |
| F | | | | | | |



# The Choice of the Distance Measure

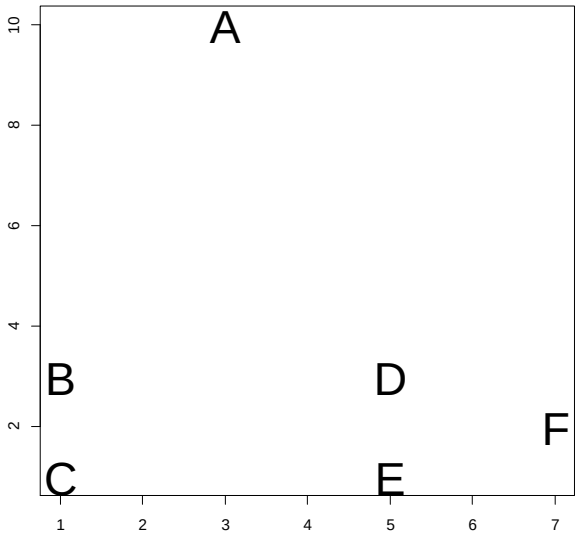
- Dissimilarity
- Similarity

# Algorithm Step 3

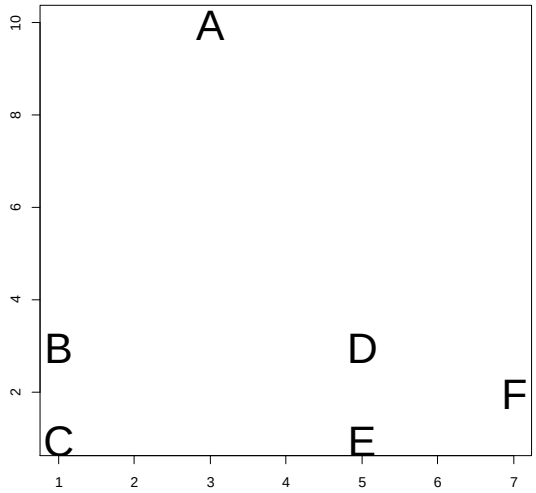
3. Find the closest (most similar) pair of clusters and merge them into a single cluster.

| ID | x | y |
|---|---|---|
| A | 3 | 10 |
| B | 1 | 3 |
| C | 1 | 1 |
| D | 5 | 3 |
| E | 5 | 1 |
| F | 7 | 2 |

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | | | | | | |
| B | | | | | | |
| C | | | | | | |
| D | | | | | | |
| E | | | | | | |
| F | | | | | | |



# Algorithm Step 4



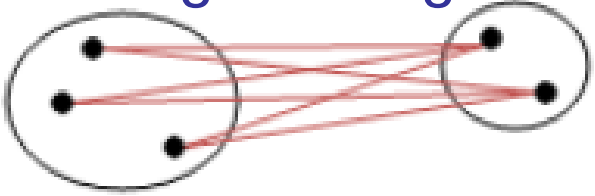

| ID | x | y |
|---|---|---|
| A | 3 | 10 |
| B | 1 | 3 |
| C | 1 | 1 |
| D | 5 | 3 |
| E | 5 | 1 |
| F | 7 | 2 |

4. Repeat 2. and 3. until all items are merged into a single cluster.

| | A | BC | DE | F |
|---|---|---|---|---|
| A | | | | |
| BC | | | | |
| DE | | | | |
| F | | | | |

# The Linkages

## Average Linkage

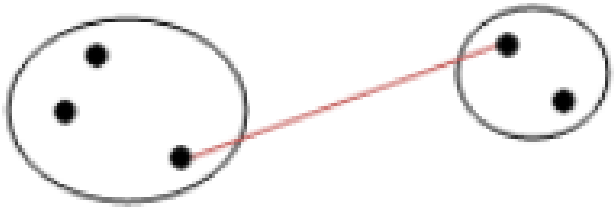


The average distance between all pairs of observations in the two clusters.

## Complete Linkage

The maximum of the distances

## Single Linkage



The minimum of the distances

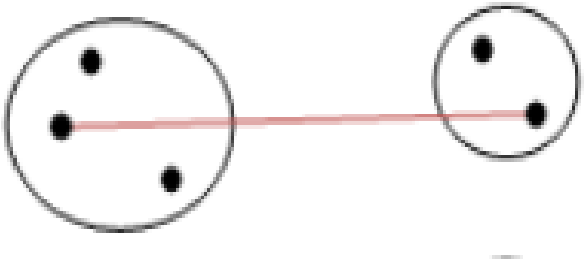
## Centroid Linkage

The distance between centers of two clusters.

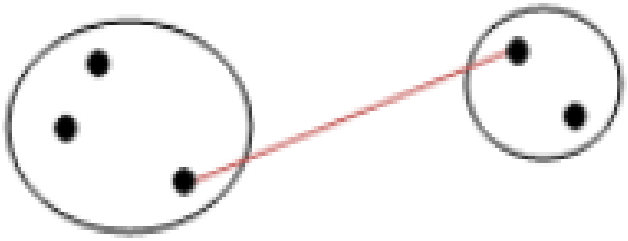
# Wald's Linkage

# Linkage Comparison

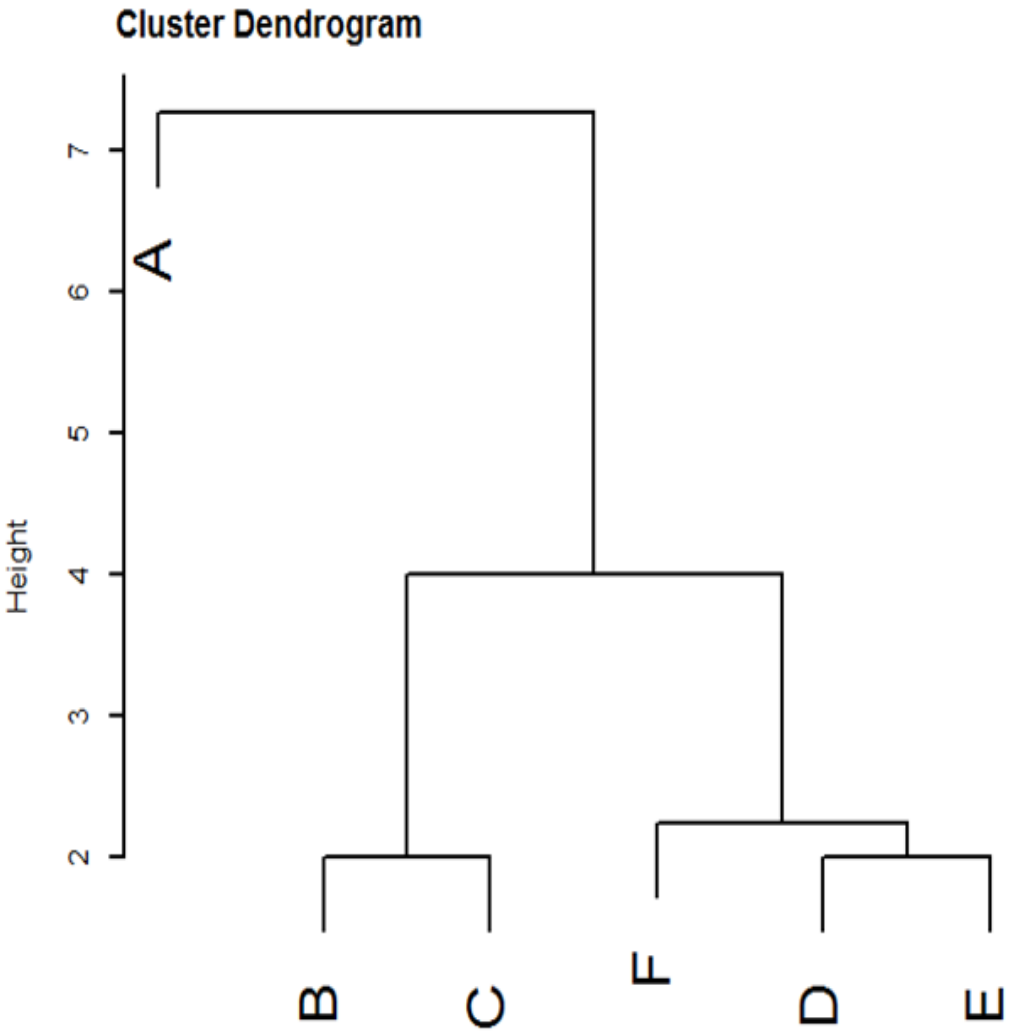
## Complete Linkage



## Single Linkage



# Clustering Result using Euclidean Distance with Single Linkage




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | | | | | | |
| B | | | | | | |
| C | | | | | | |
| D | | | | | | |
| E | | | | | | |
| F | | | | | | |

| | A | BC | DE | F |
|---|---|---|---|---|
| A | | | | |
| BC | | | | |
| DE | | | | |
| F | | | | |

# Strength and Weakness

- Strength
  Do not require a prior knowledge

- Weakness
  A mistake cannot be undone

- Computation Complexity $O(n^3)$

# Votes for Republican Candidate

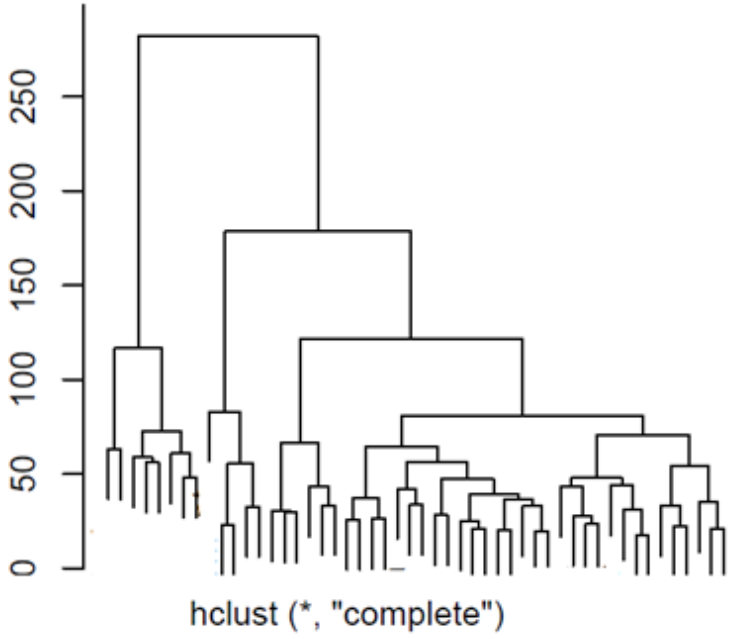
| | 1856 | 1860 | 1864 | 1868 | 1872 | 1876 | ≈ | 1968 | 1972 | 1976 |
|---|---|---|---|---|---|---|---|---|---|---|
| Alabama | NA | NA | NA | 51.44 | 53.19 | 40.02 | | 14 | 72.4 | 43.48 |
| Alaska | NA | NA | NA | NA | NA | NA | | 45.3 | 58.1 | 62.91 |
| Arizona | NA | NA | NA | NA | NA | NA | | 54.8 | 64.7 | 58.62 |
| Arkansas | NA | NA | NA | 53.73 | 52.17 | 39.88 | | 30.8 | 68.9 | 34.97 |
| California | 18.77 | 32.96 | 58.63 | 50.24 | 56.38 | 50.88 | | 47.8 | 55 | 50.89 |
| Colorado | NA | NA | NA | NA | NA | NA | | 50.5 | 62.6 | 55.89 |
| Connecticut | 53.18 | 53.86 | 51.38 | 51.54 | 52.25 | 48.34 | | 44.3 | 58.6 | 52.64 |
| Delaware | 2.11 | 23.71 | 48.2 | 40.98 | 50.99 | 44.55 | | 45.1 | 59.6 | 47.27 |
| Florida | NA | NA | NA | NA | 53.52 | 50.99 | | 40.5 | 71.9 | 46.83 |
| Georgia | NA | NA | NA | 35.72 | 43.77 | 27.94 | | 30.4 | 75 | 33.02 |
| Hawaii | NA | NA | NA | NA | NA | NA | | 38.7 | 62.5 | 48.72 |
| Idaho | NA | NA | NA | NA | NA | NA | | 56.8 | 64.2 | 61.77 |
| Illinois | 40.25 | 50.68 | 54.41 | 55.69 | 56.27 | 50.09 | | 47.1 | 59 | 51.11 |
| Indiana | 40.03 | 51.09 | 53.6 | 51.39 | 53 | 48.27 | | 50.3 | 66.1 | 53.77 |
| Iowa | 49.13 | 54.87 | 64.23 | 61.92 | 64.18 | 58.58 | | 53 | 57.6 | 50.51 |
| Kansas | NA | NA | 78.61 | 68.89 | 66.64 | 63.1 | | 54.8 | 67.7 | 53.91 |
| Kentucky | 0.26 | 0.93 | 30.17 | 25.45 | 46.45 | 37.61 | | 43.8 | 63.4 | 46.24 |
| Louisiana | NA | NA | NA | 29.31 | 55.69 | 51.57 | | 23.5 | 65.3 | 47 |
| Maine | 61.37 | 64.15 | 60.22 | 62.42 | 67.86 | 56.73 | | 43.1 | 61.5 | 50.34 |
| Maryland | 0.32 | 3.11 | 55.1 | 32.8 | 49.65 | 43.94 | | 41.9 | 61.3 | 46.87 |
| Massachusetts | 64.72 | 62.75 | 72.22 | 69.67 | 69.25 | 57.74 | | 32.9 | 45.2 | 41.93 |
| Michigan | 56.98 | 57.18 | 55.89 | 56.98 | 62.67 | 52.45 | | 41.5 | 56.2 | 52.68 |
| Minnesota | NA | 63.42 | 59.06 | 60.8 | 61.55 | 58.77 | | 41.5 | 51.6 | 44.3 |

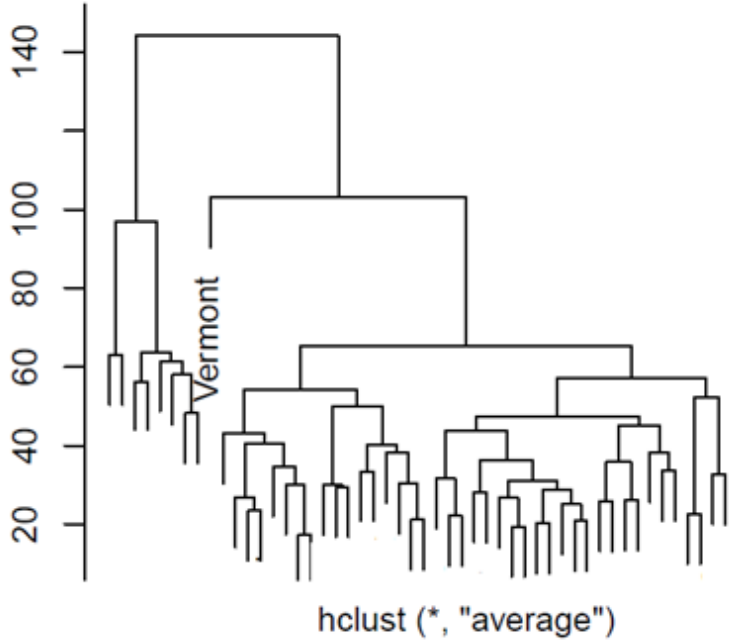
# Hierarchical Clustering Results

```
votes=votes.repub
years=as.numeric(gsub("X", "", colnames(votes.repub)))
names(votes)=years

aa1=hclust(dist(votes), method="complete")
aa2=hclust(dist(votes), method="average")
par(mfrow=c(1,2))
plot(aa1)
plot(aa2)
```

```
library(cluster)
library(dendextend)
library(usmap)
library(gplots)
```



# Hierarchical Clustering Results

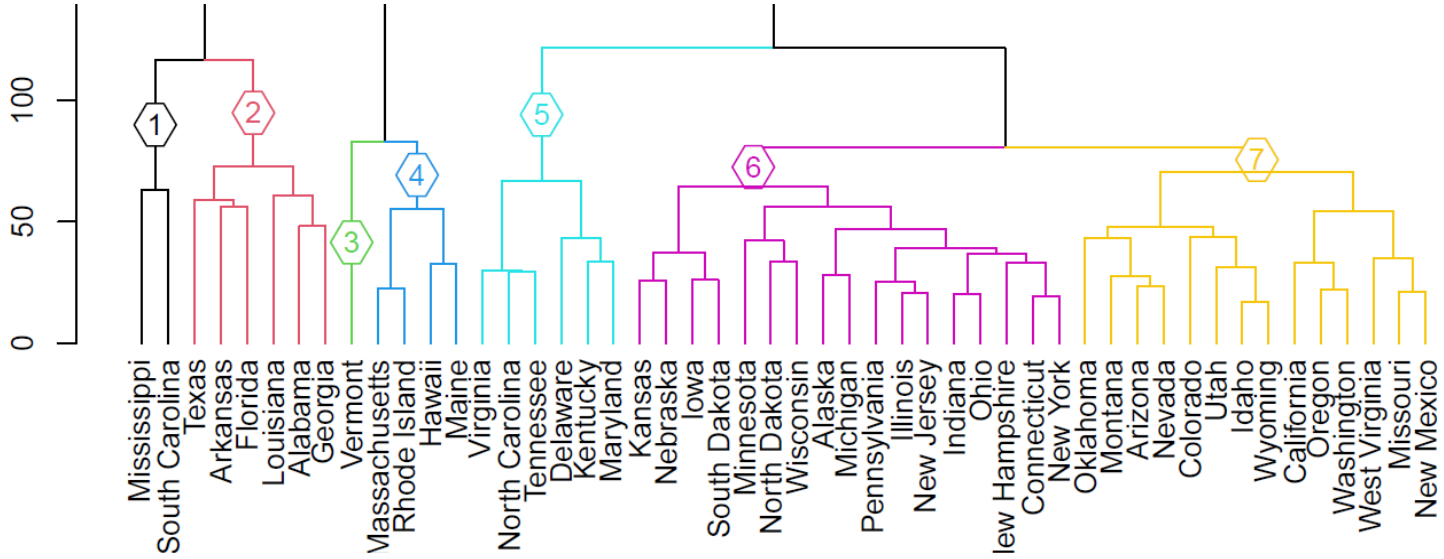
```
cluster=cutree(aa1, h=75, order_clusters_as_data=FALSE)
cluster[1:5]
```

```
##    Mississippi South Carolina    Texas     Arkansas    Florida
##              1              1        2            2          2
```

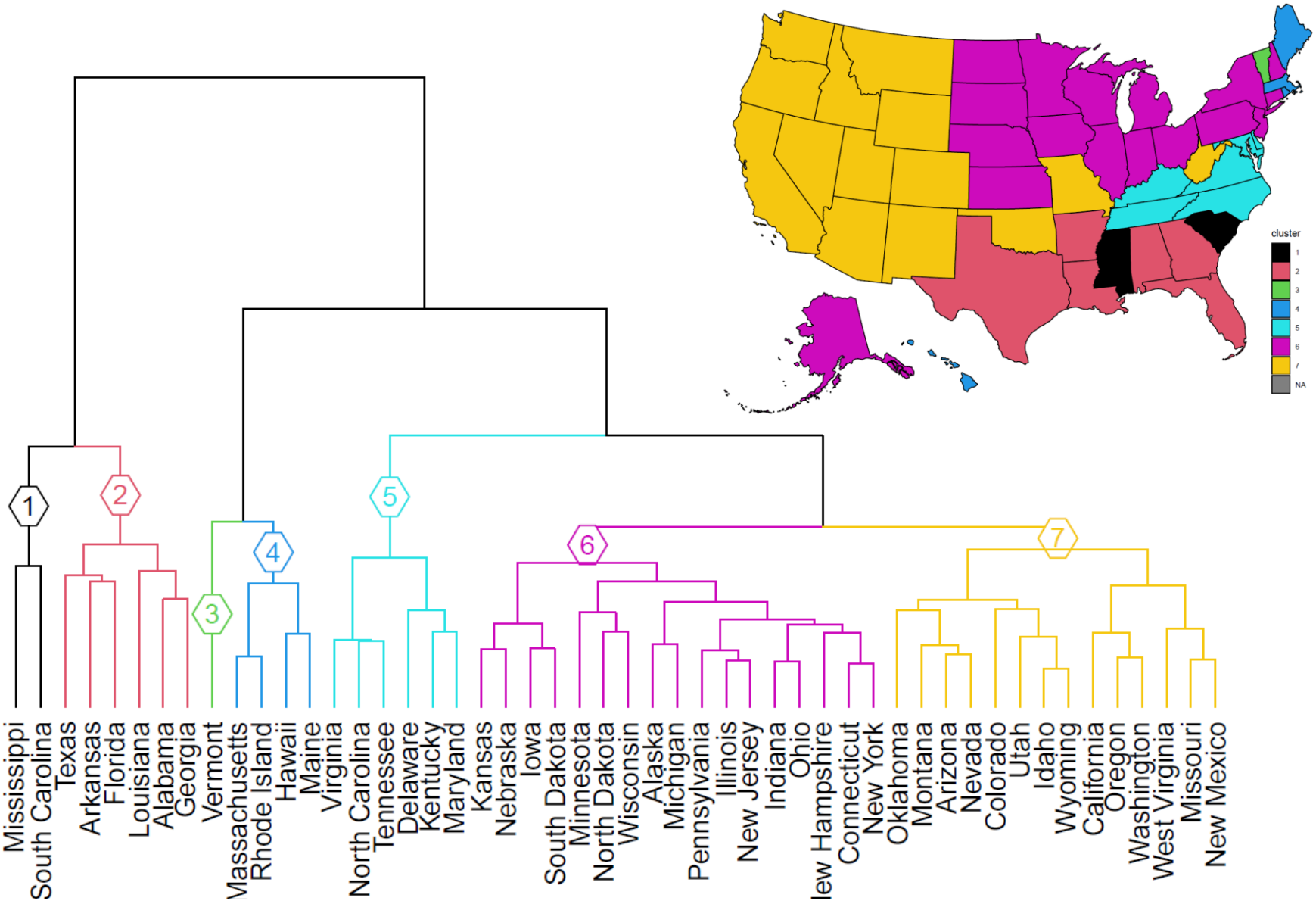
```
table(cluster)
```

```
## cluster
##  1  2  3  4  5  6  7
##  2  6  1  4  6 17 14
```

```
aa1col=color_branches(as.dendrogram(aa1), h=75, col=1:7, groupLabels=TRUE)
plot(aa1col)
```

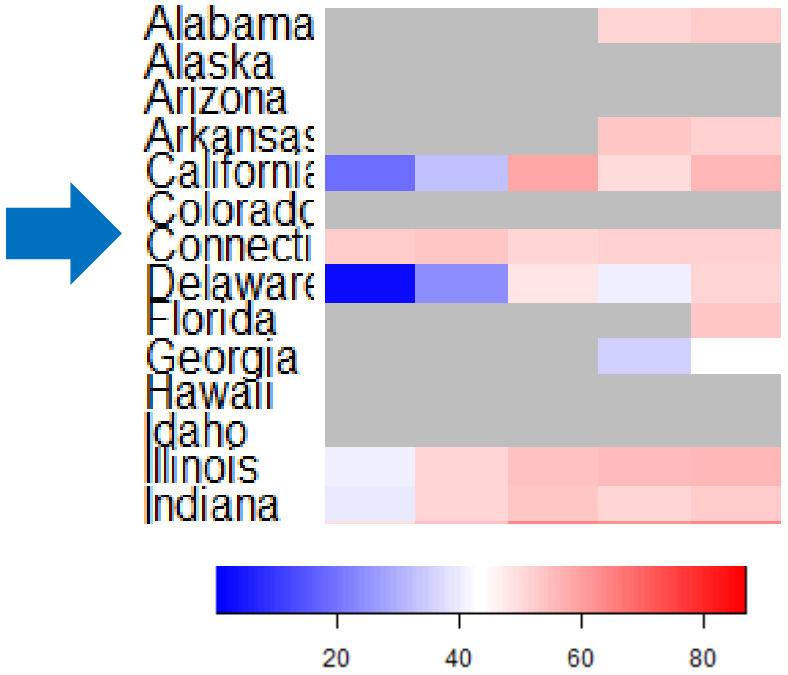


# Hierarchical Clustering Results



# Heatmap – Visualize Patterns

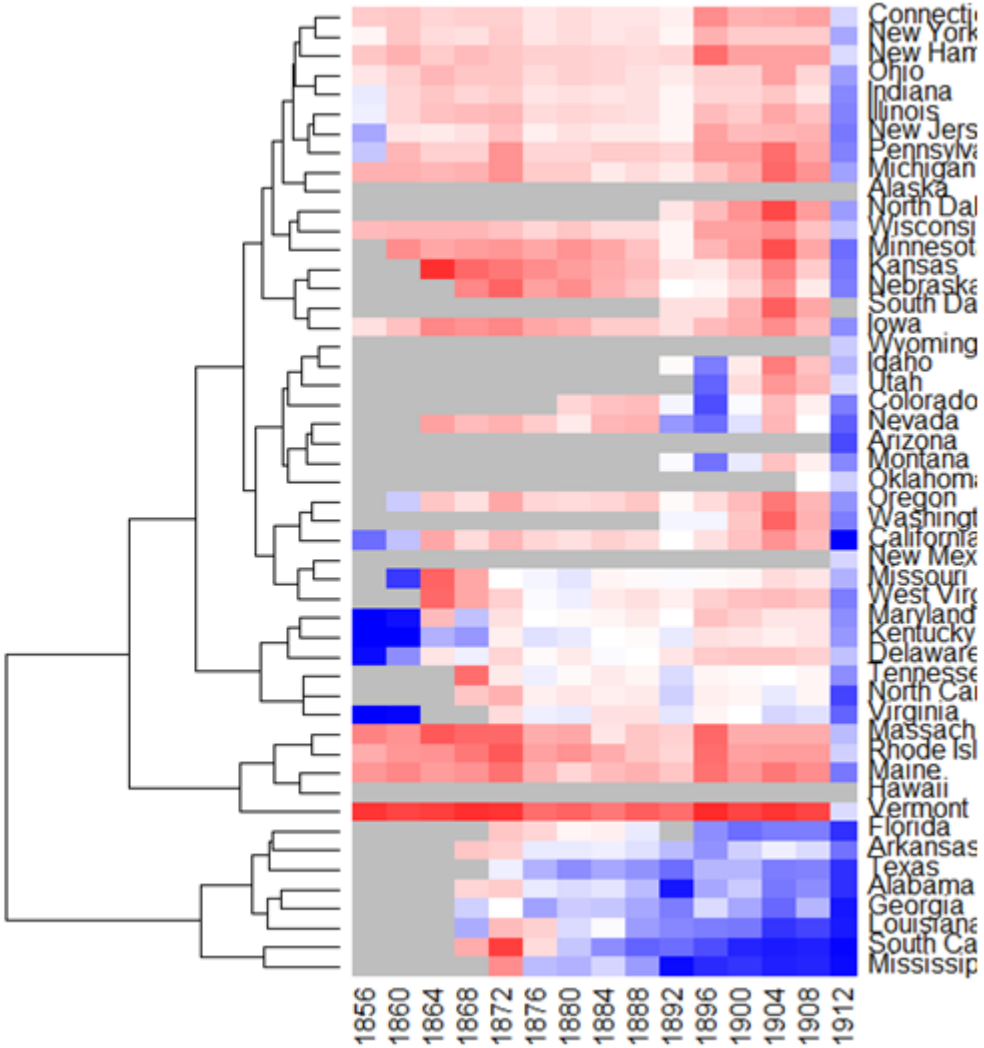
| | 1856 | 1860 | 1864 | 1868 |
|---|---|---|---|---|
| Alabama | NA | NA | NA | 51.44 |
| Alaska | NA | NA | NA | NA |
| Arizona | NA | NA | NA | NA |
| Arkansas | NA | NA | NA | 53.73 |
| California | 18.77 | 32.96 | 58.63 | 50.24 |
| Colorado | NA | NA | NA | NA |
| Connecticut | 53.18 | 53.86 | 51.38 | 51.54 |
| Delaware | 2.11 | 23.71 | 48.2 | 40.98 |
| Florida | NA | NA | NA | NA |
| Georgia | NA | NA | NA | 35.72 |
| Hawaii | NA | NA | NA | NA |
| Idaho | NA | NA | NA | NA |
| Illinois | 40.25 | 50.68 | 54.41 | 55.69 |
| Indiana | 40.03 | 51.09 | 53.6 | 51.39 |



# Data Re-ordering

## By Original Order (by States)


## By Clustering Results



# Cluster Representative Features

```
hc=function(x) hclust(x, method="complete")
dd=function(x) dist(x, method="euclidean")
```

```
heatmap.2(
 as.matrix(votes),
 Rowv = aa1col,
 Colv = NULL,
 distfun=dd,
 hclustfun=hc,
 dendrogram="row",
 col=bluered(100),
 trace="none",
 density="none",
 na.color="grey",
 keysize=1.2,
 cexCol=1,
 cexRow=0.8
)
```