Unsupervised Learning

Vahid Partovi Nia

Lecture 10

McGill

Outline

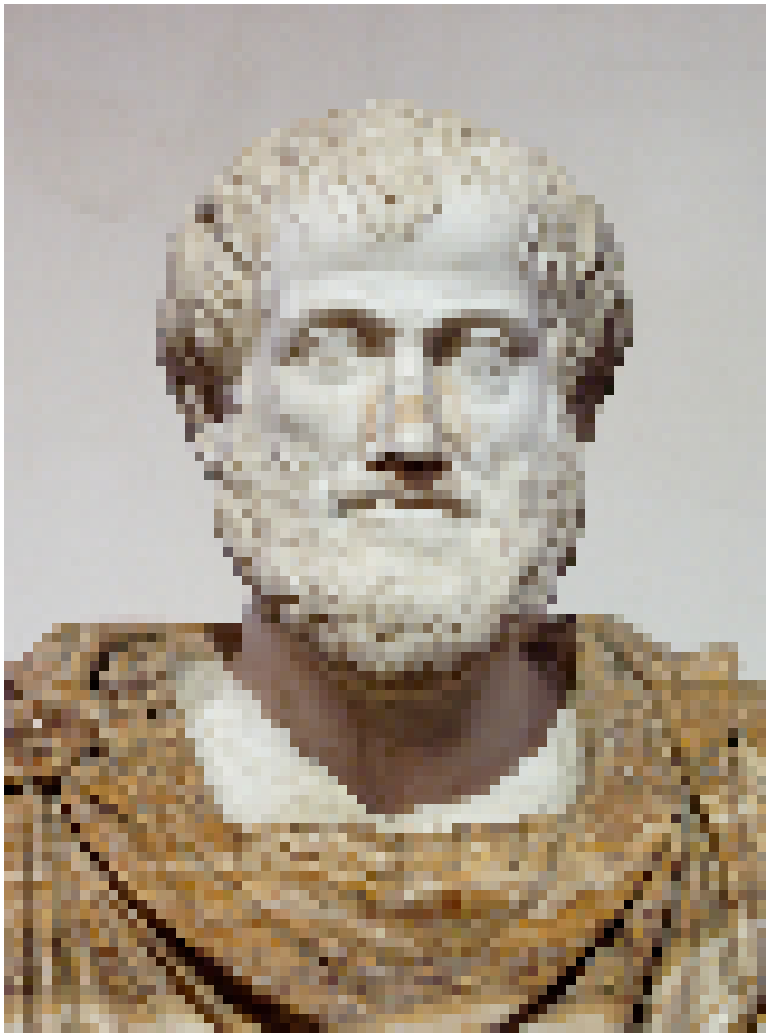
1. Introduction
2. K-Means
3. Hierarchical Clustering

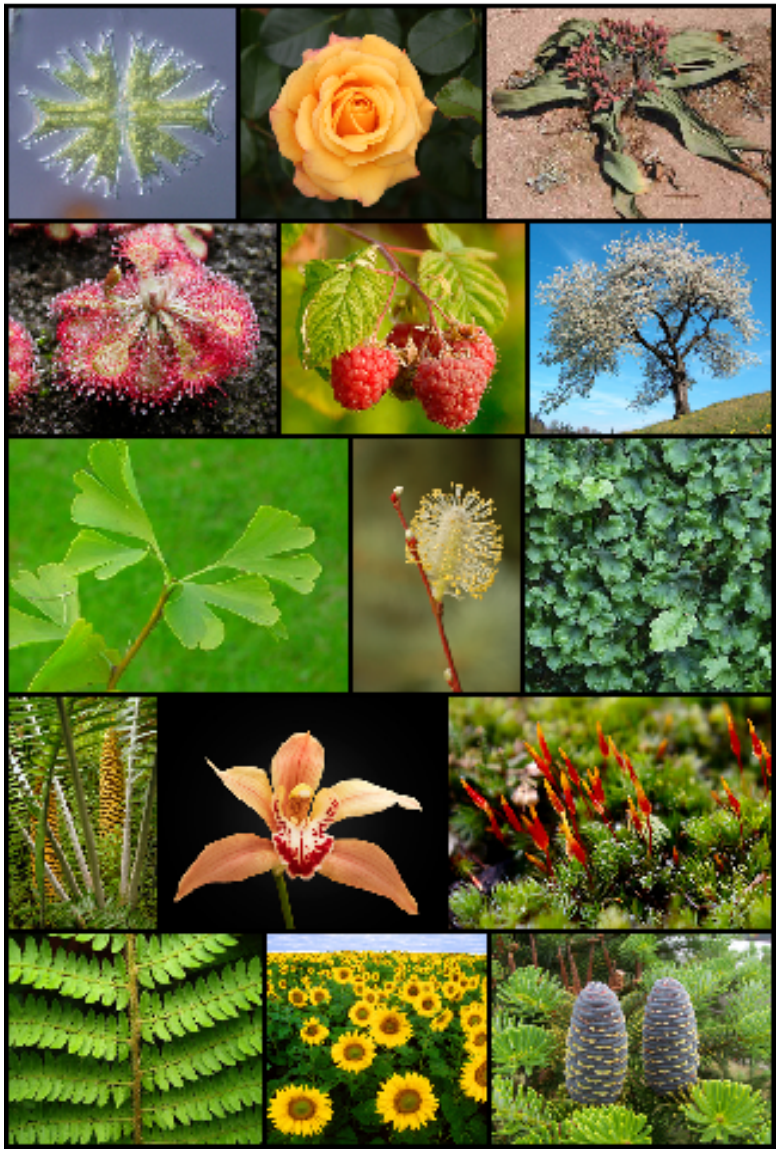
Early Learning

Introduction

K-Means

Hierarchical Clustering





Avicenna

Introduction

K-Means

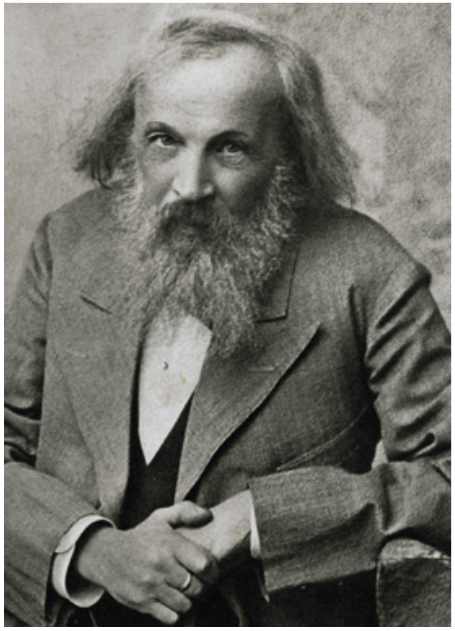
Hierarchical Clustering

Mendeleev

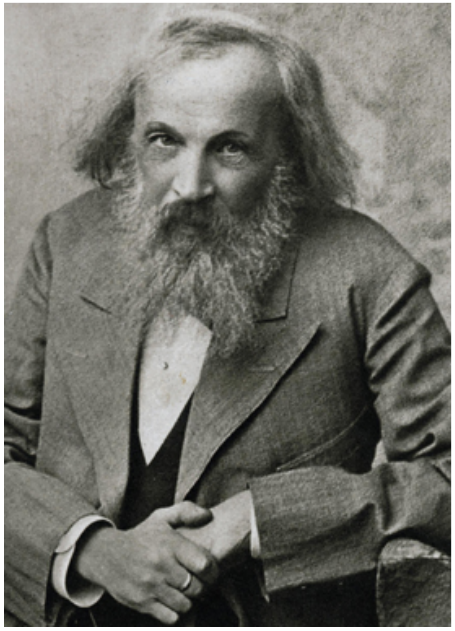
Introduction

K-Means

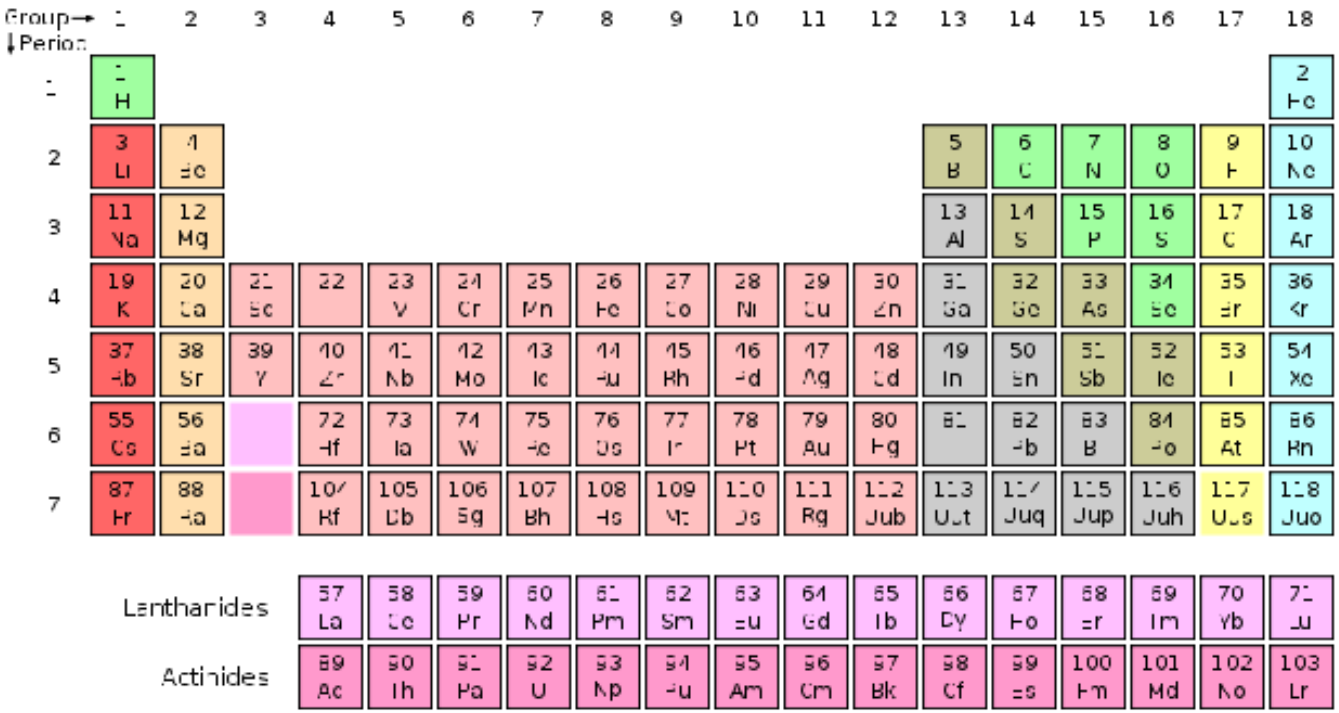
Hierarchical Clustering

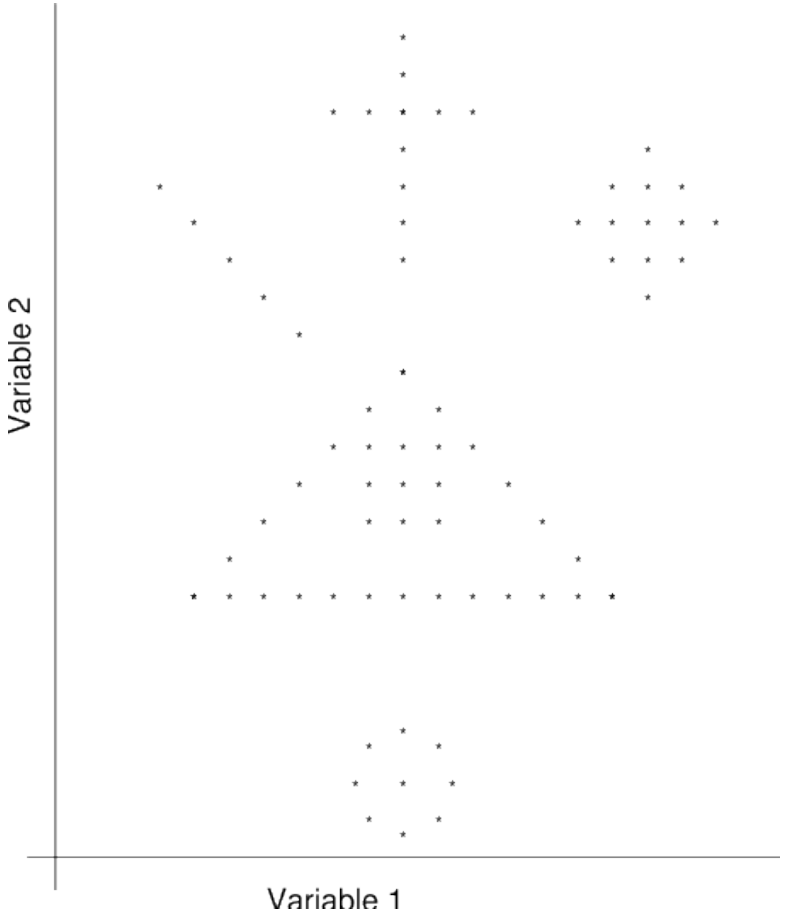


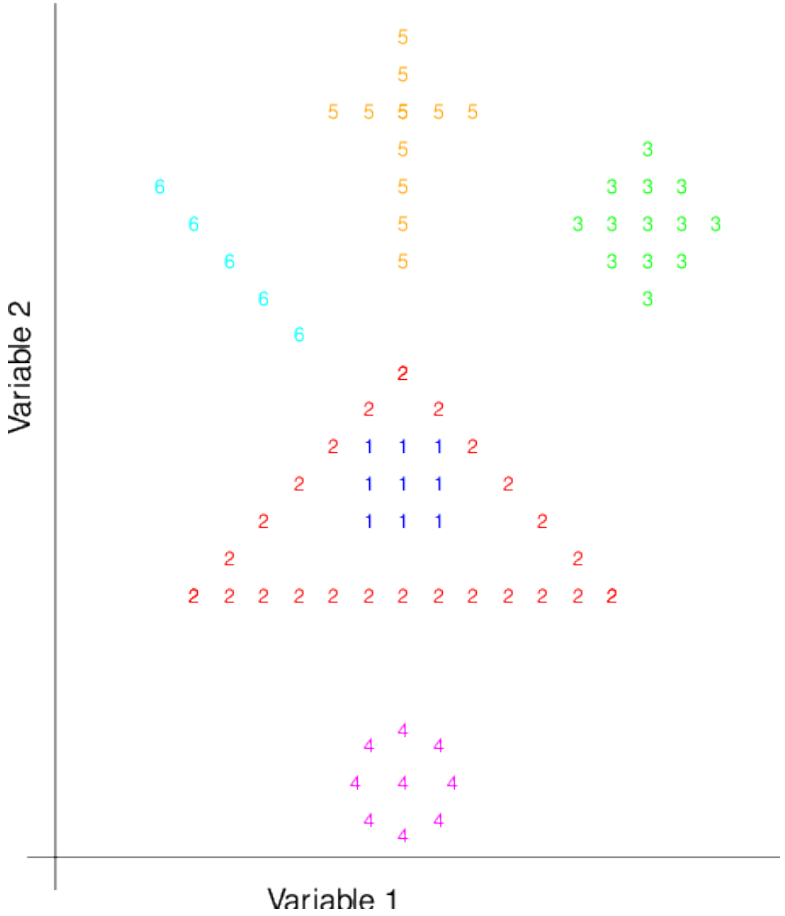
Mendeleev



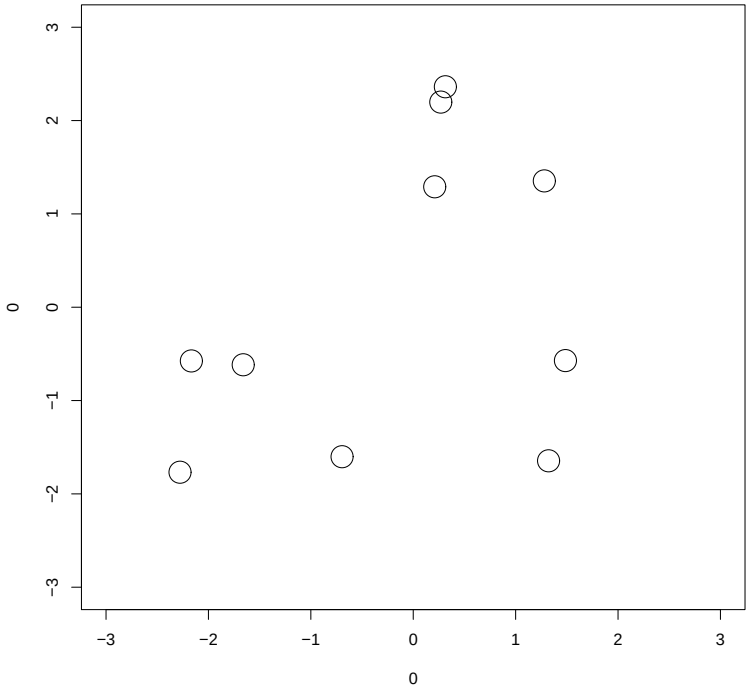
I	II	III	IV	V	VI	VII			VIII
H 1.01									
Li 6.94	Be 9.01	B 10.8	C 12.0	N 14.0	O 16.0	F 19.0			
Na 23.0	Mg 24.3	Al 27.0	Si 28.1	P 31.0	S 32.1	Cl 35.5			
K 39.1	Ca 40.1		Ti 47.9	V 50.9	Cr 52.0	Mn 54.9	Fe 55.9	Co 58.9	Ni 58.7
Cu 63.5	Zn 65.4			As 74.9	Se 79.0	Br 79.9			
Rb 85.5	Sr 87.6	Y 88.9	Zr 91.2	Nb 92.9	Mo 95.9		Ru 101	Rh 103	Pd 106
Ag 108	Cd 112	In 115	Sn 119	Sb 122	Te 128	I 127			
Ce 133	Ba 137	La 139		Ta 181	W 184		Os 194	Ir 192	Pt 195
Au 197	Hg 201	Ti 204	Pb 207	Bi 209					
			Th 232		U 238				

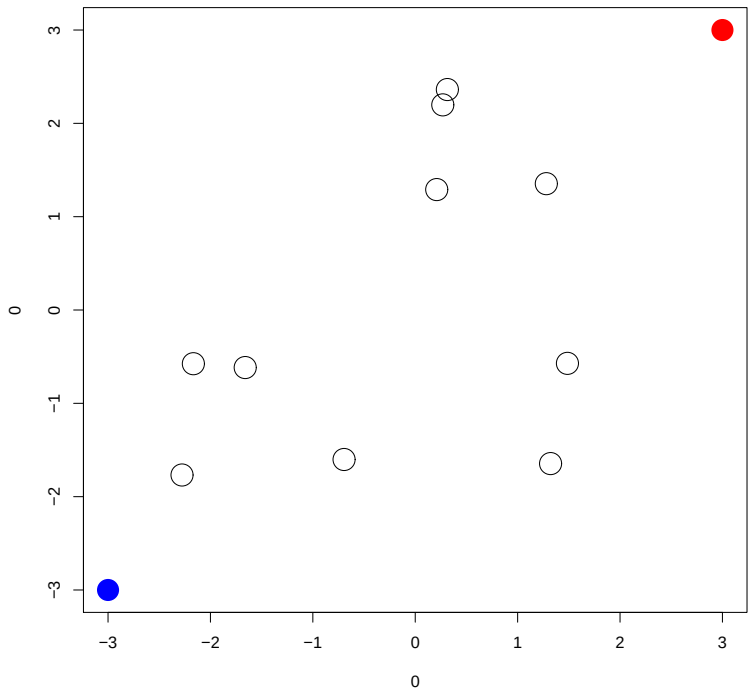

Group→
Period↓
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
1 H
2 He
2 Li Be B C N O F Ne
3 Na Mg Al Si P S Cl Ar
4 K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
5 Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
6 Cs Ba Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn
7 Fr Ra Rf Db Sg Bh Hs Mt Ds Rg Cn Nh Fl Mc Lv Ts Og
Lanthanides La Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu
Actinides Ac Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No Lr

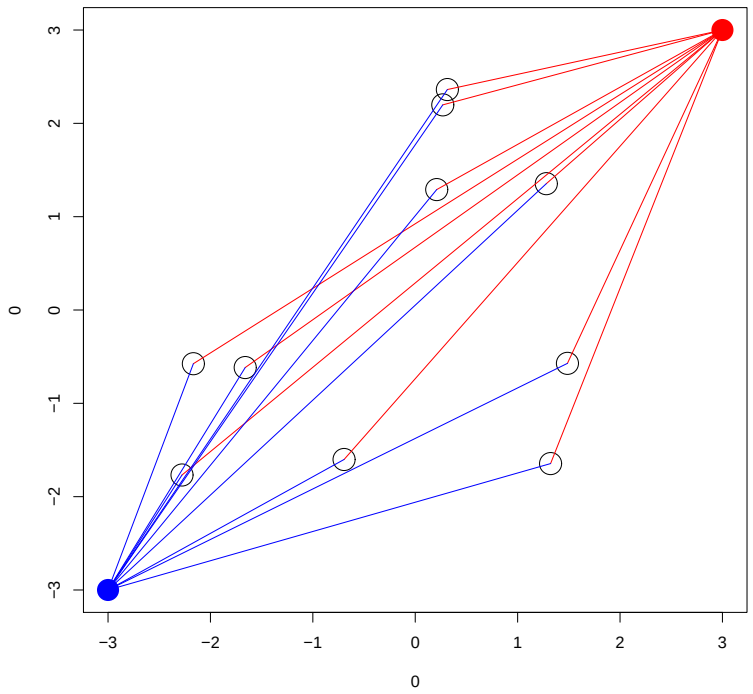

Variable 2
Variable 1

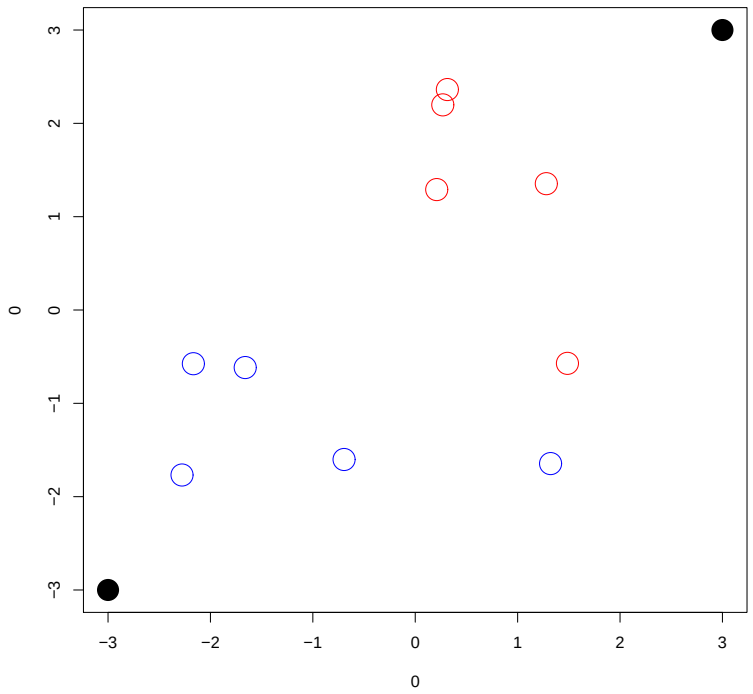

Variable 2
Variable 1

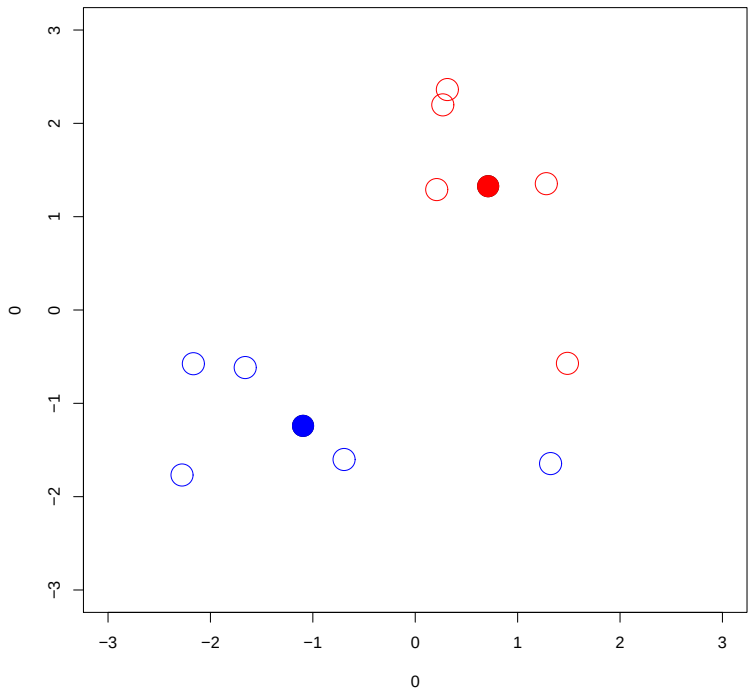
Initial 1

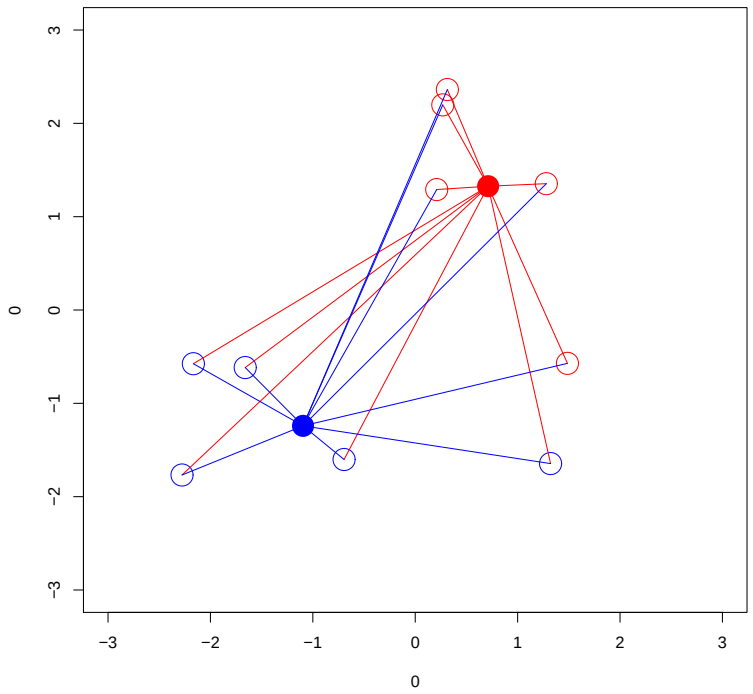


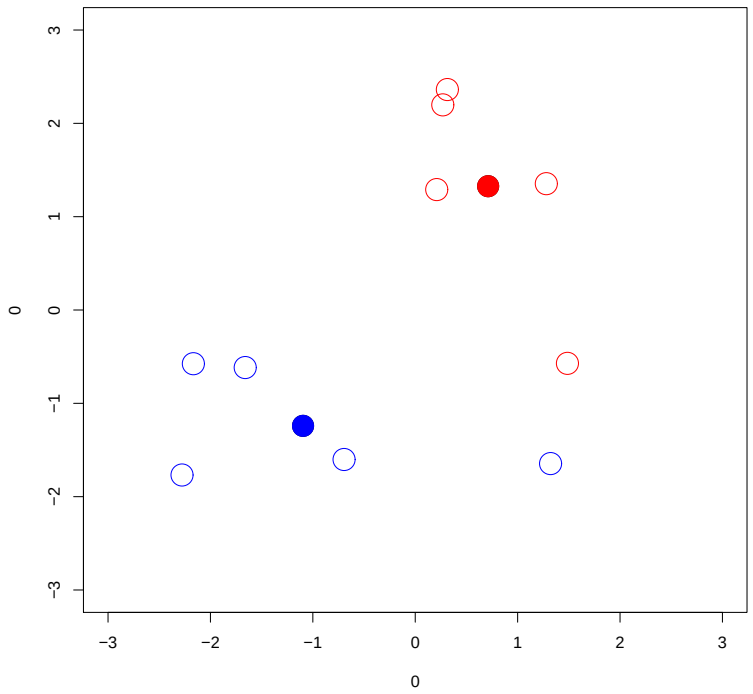

3
2
1
0
−1
−2
−3
0
−3
−2
−1
0
1
2
3
0


3
2
1
0
−1
−2
−3
0
−3
−2
−1
0
1
2
3
0

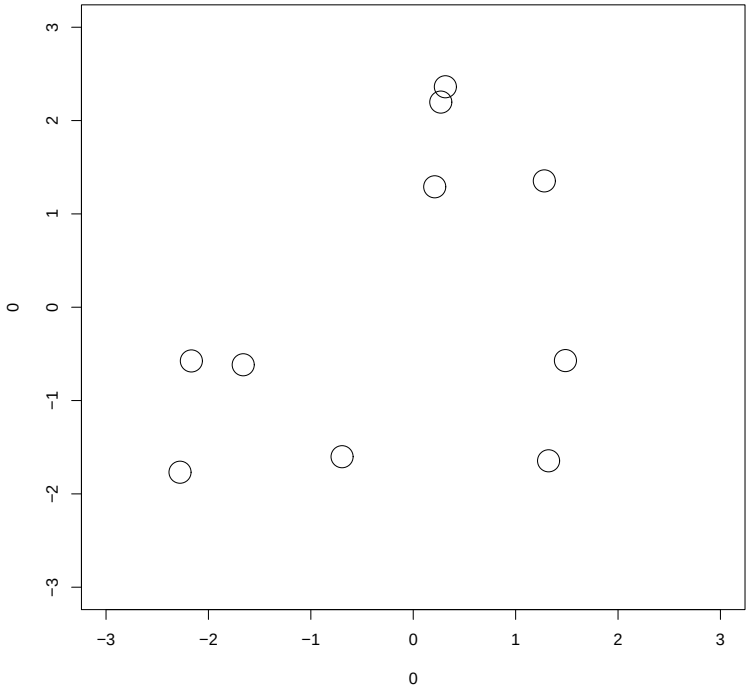

3
2
1
0
−1
−2
−3
0
−3
−2
−1
0
1
2
3
0

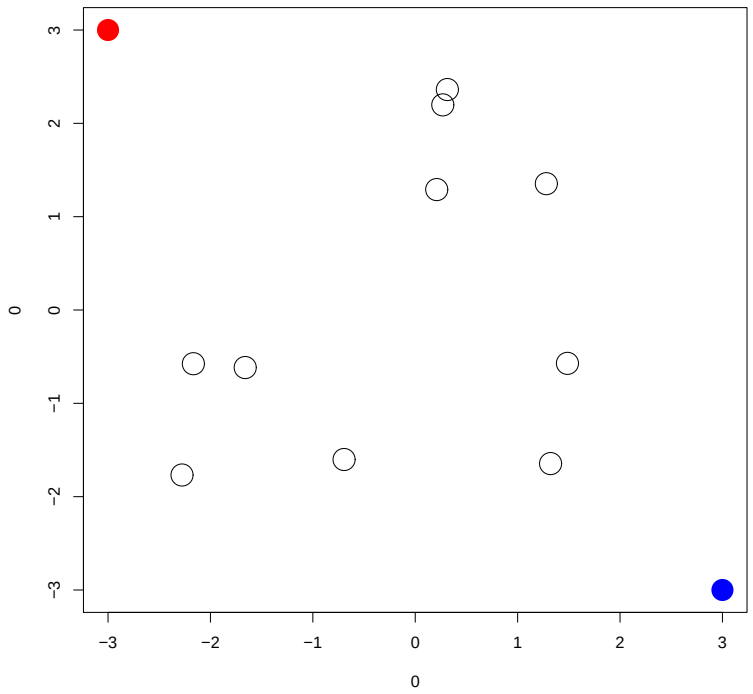

3
2
1
0
−1
−2
−3
0
−3
−2
−1
0
1
2
3
0

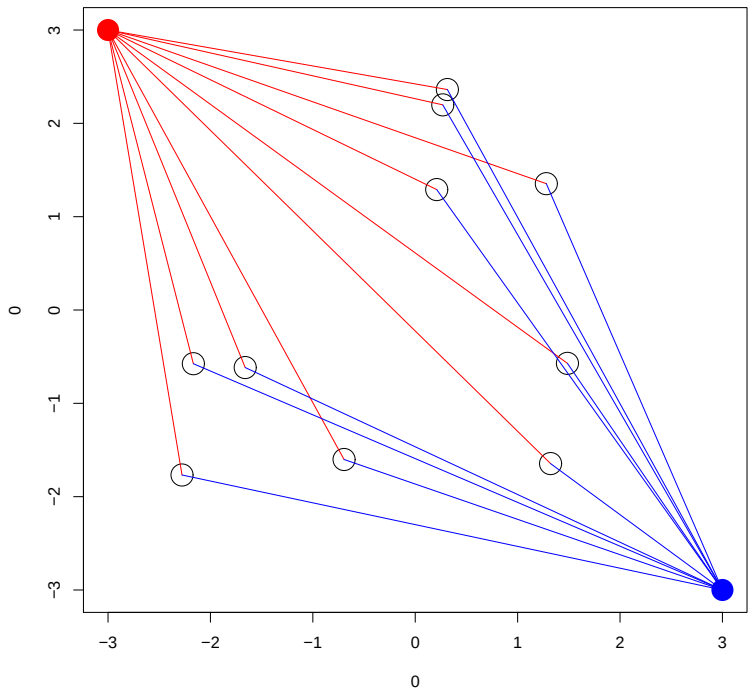

3
2
1
0
−1
−2
−3
0
−3
−2
−1
0
1
2
3
0

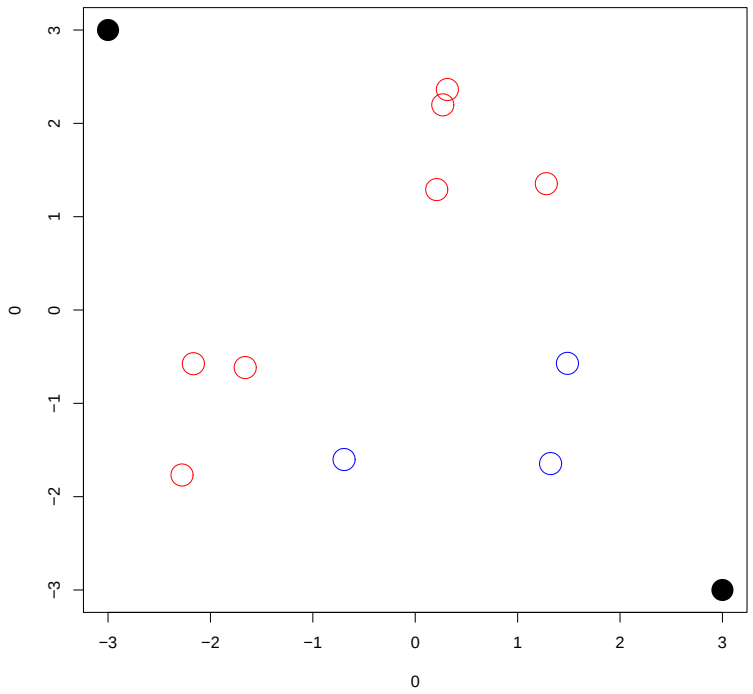

3
2
1
0
−1
−2
−3
0
−3
−2
−1
0
1
2
3
0

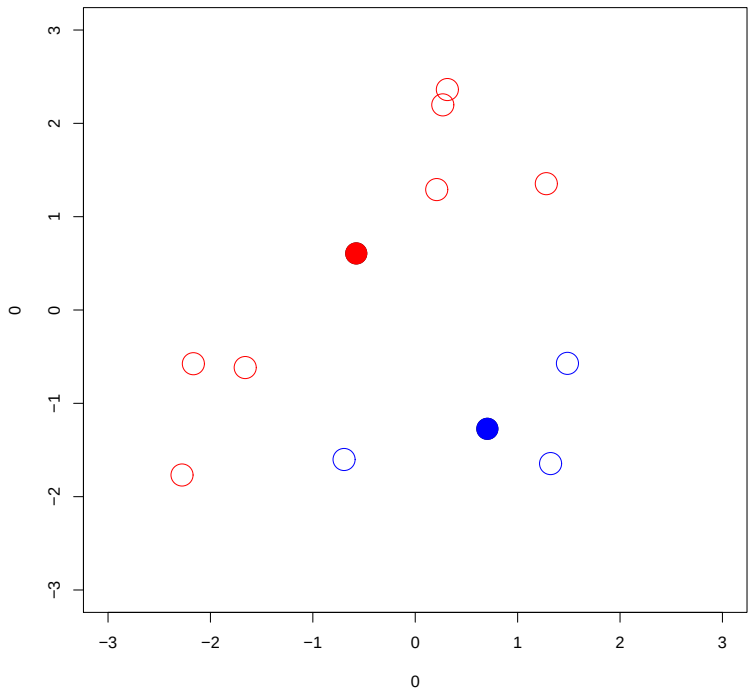
Initial 2

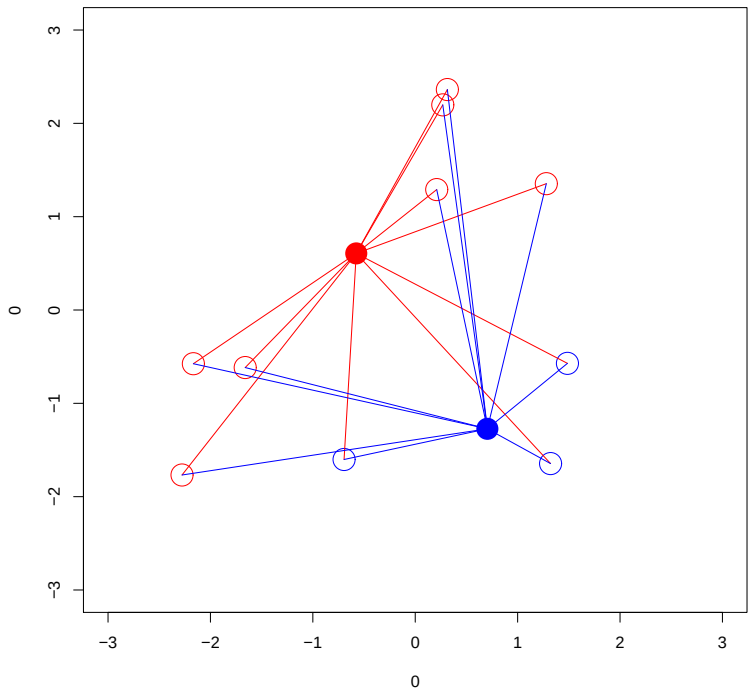


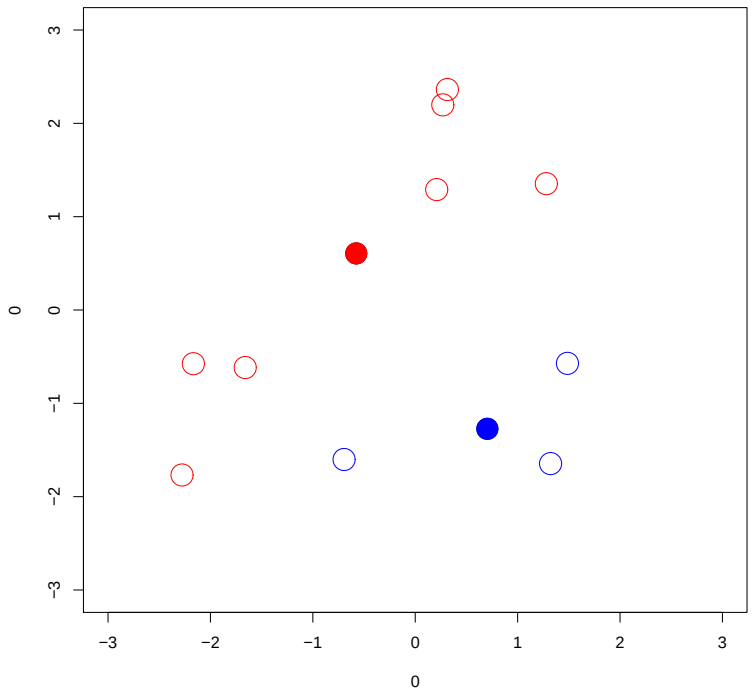

3
2
1
0
−1
−2
−3
0
−3
−2
−1
0
1
2
3
0


3
2
1
0
−1
−2
−3
0
−3
−2
−1
0
1
2
3
0

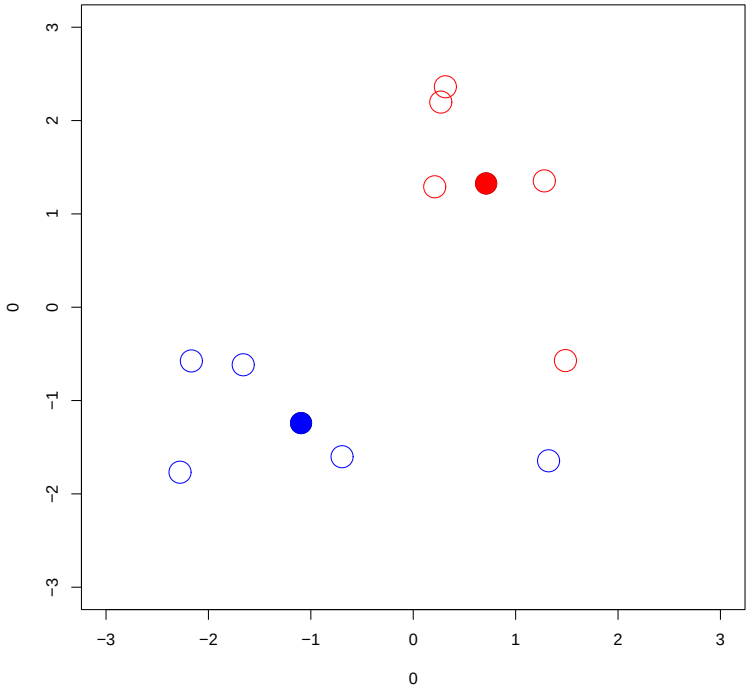

3
2
1
0
−1
−2
−3
0
−3
−2
−1
0
1
2
3
0


3
2
1
0
−1
−2
−3
0
−3
−2
−1
0
1
2
3
0

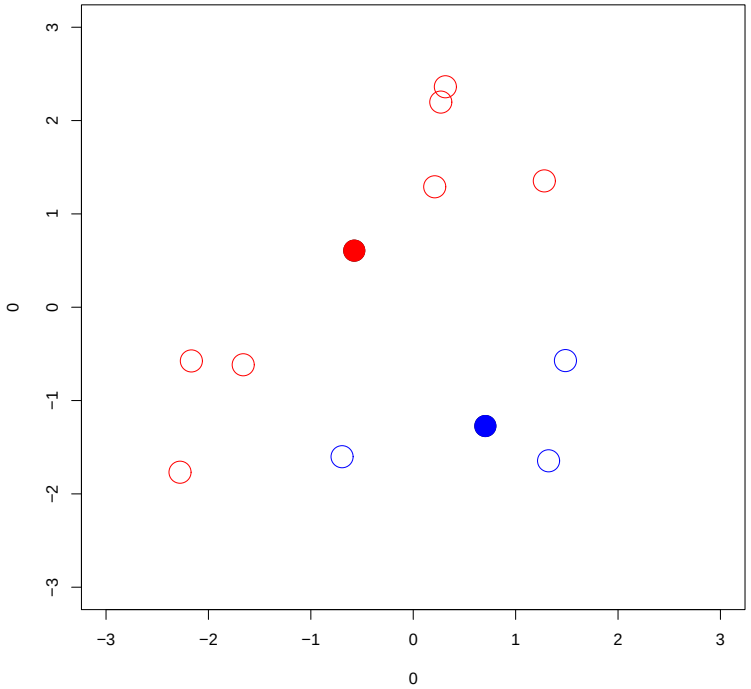

3
2
1
0
−1
−2
−3
0
−3
−2
−1
0
1
2
3
0


3
2
1
0
−1
−2
−3
0
−3
−2
−1
0
1
2
3
0

Compare 1



Compare 2



Formalism

$$S_T = S_W(\mathbf{c}) + S_B(\mathbf{c}), \tag{1}$$

in which

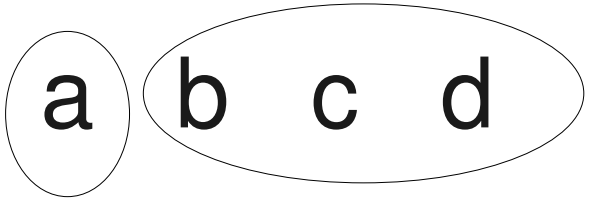
$$S_W(\mathbf{c}) = \sum_{c=1}^{k} \sum_{\{i|d_i=c\}} \sum_{\{i'|d_{i'}=c\}} (y_{ci} - \bar{y}_c)^2, \tag{2}$$

$$S_B(\mathbf{c}) = \sum_{c=1}^{k} \sum_{\{i|d_i=c\}} \sum_{\{i'|d_{i'}\neq c\}} (y_{ci} - \bar{y}_c)^2 \tag{3}$$

Grouping Data

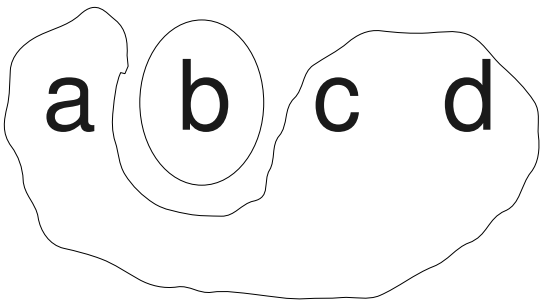
{a,b,c,d}

{a}{b,c,d}



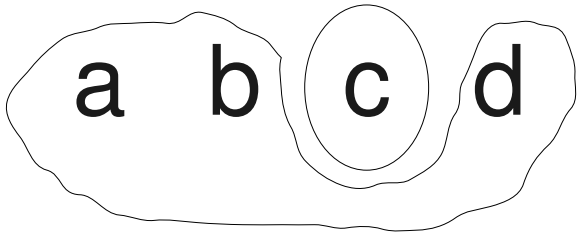
{a,b,c,d}

{a}{b,c,d} {b}{a,c,d}



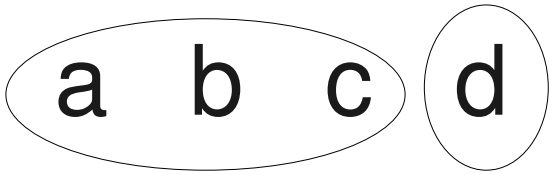
{a,b,c,d}

{a}{b,c,d} {b}{a,c,d} {c}{a,b,d}



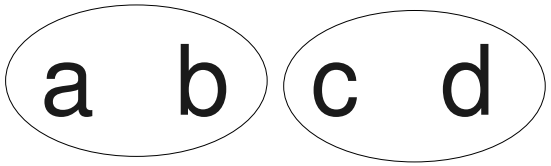
{a,b,c,d}

{a}{b,c,d} {b}{a,c,d} {c}{a,b,d} {d}{a,b,c}



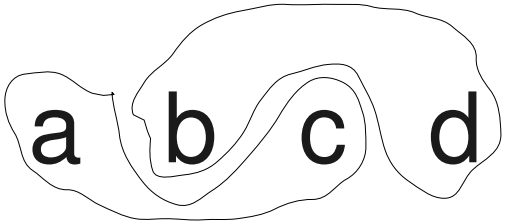
{a,b,c,d}

{a}{b,c,d} {b}{a,c,d} {c}{a,b,d} {d}{a,b,c} {a,b}{c,d}



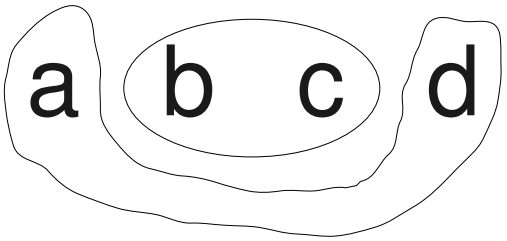
{a,b,c,d}

{a}{b,c,d} {b}{a,c,d} {c}{a,b,d} {d}{a,b,c} {a,b}{c,d} {a,c}{b,d}



{a,b,c,d}

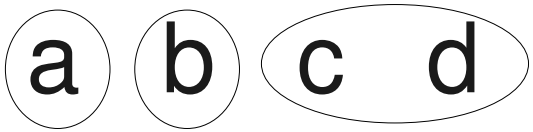
{a}{b,c,d} {b}{a,c,d} {c}{a,b,d} {d}{a,b,c} {a,b}{c,d} {a,c}{b,d} {a,d}{b,c}



{a,b,c,d}

{a}{b,c,d} {b}{a,c,d} {c}{a,b,d} {d}{a,b,c} {a,b}{c,d} {a,c}{b,d} {a,d}{b,c}

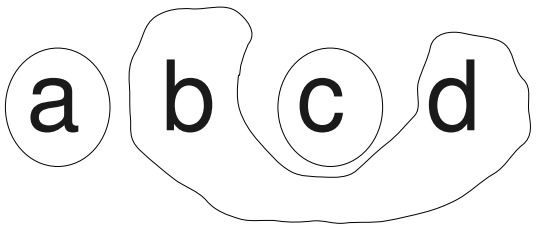
{a},{b}{c,d}



{a,b,c,d}

{a}{b,c,d} {b}{a,c,d} {c}{a,b,d} {d}{a,b,c} {a,b}{c,d} {a,c}{b,d} {a,d}{b,c}

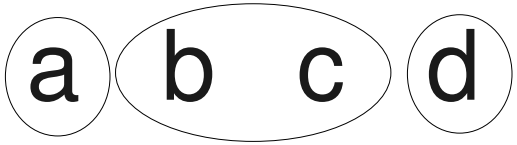
{a},{b}{c,d} {a},{c}{b,d}



{a,b,c,d}

{a}{b,c,d} {b}{a,c,d} {c}{a,b,d} {d}{a,b,c} {a,b}{c,d} {a,c}{b,d} {a,d}{b,c}

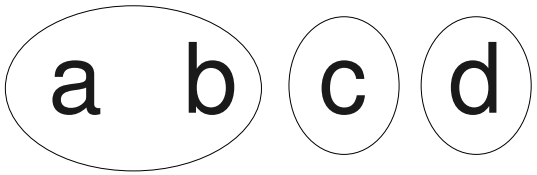
{a},{b}{c,d} {a},{c}{b,d} {a},{d}{b,c}



{a,b,c,d}

{a}{b,c,d} {b}{a,c,d} {c}{a,b,d} {d}{a,b,c} {a,b}{c,d} {a,c}{b,d} {a,d}{b,c}

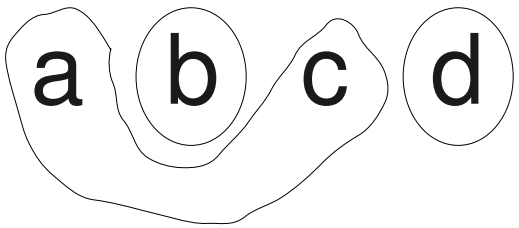
{a},{b}{c,d} {a},{c}{b,d} {a},{d}{b,c} {a,b}{c},{d}



{a,b,c,d}

{a}{b,c,d} {b}{a,c,d} {c}{a,b,d} {d}{a,b,c} {a,b}{c,d} {a,c}{b,d} {a,d}{b,c}

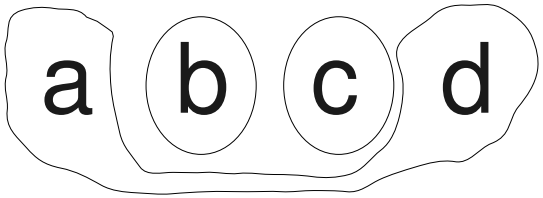
{a},{b}{c,d} {a},{c}{b,d} {a},{d}{b,c} {a,b}{c},{d} {a,c}{b},{d}



{a,b,c,d}

{a}{b,c,d} {b}{a,c,d} {c}{a,b,d} {d}{a,b,c} {a,b}{c,d} {a,c}{b,d} {a,d}{b,c}

{a},{b}{c,d} {a},{c}{b,d} {a},{d}{b,c} {a,b}{c},{d} {a,c}{b},{d} {a,d}{b},{c}

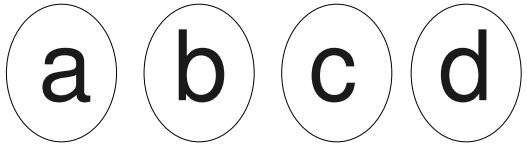


{a,b,c,d}

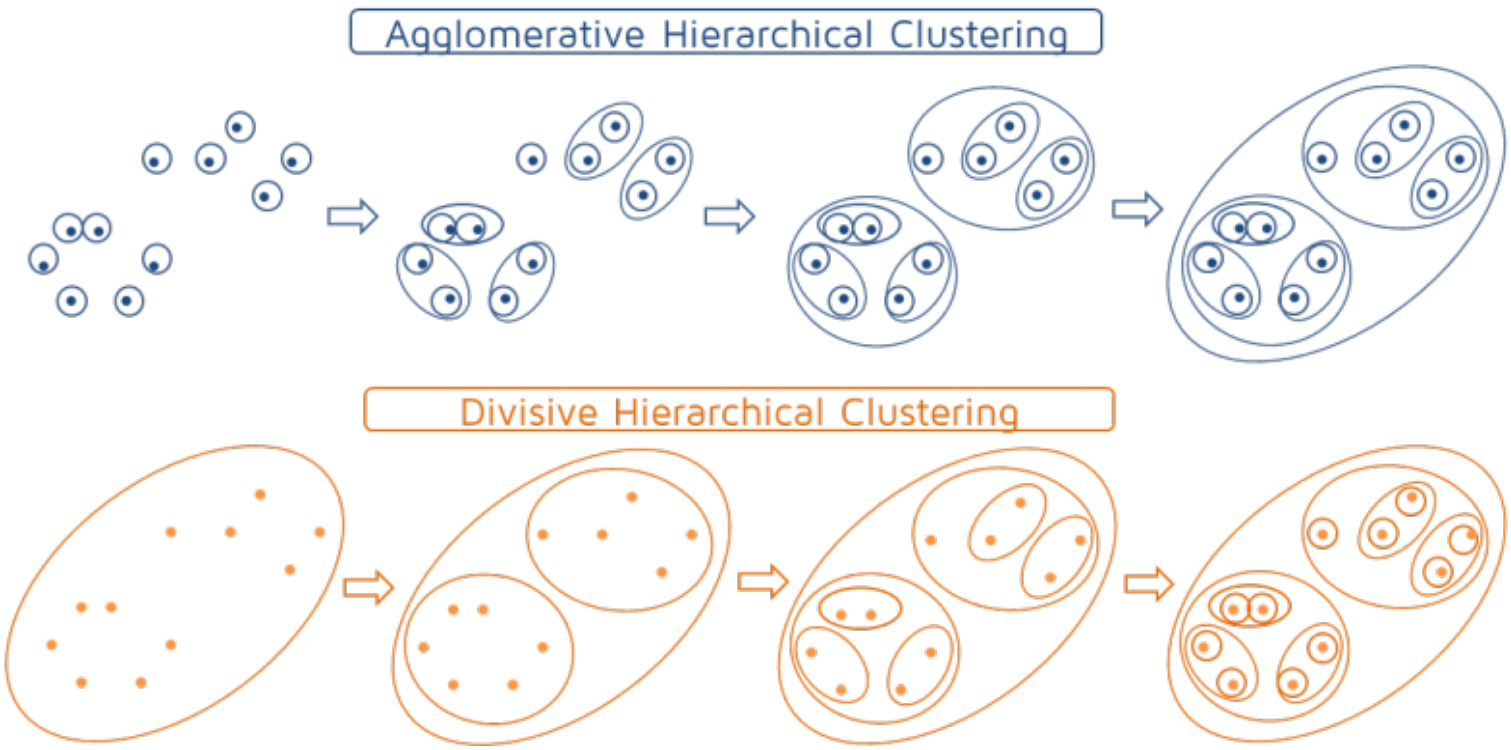
{a}{b,c,d} {b}{a,c,d} {c}{a,b,d} {d}{a,b,c} {a,b}{c,d} {a,c}{b,d} {a,d}{b,c}

{a},{b}{c,d} {a},{c}{b,d} {a},{d}{b,c} {a,b}{c},{d} {a,c}{b},{d} {a,d}{b},{c}

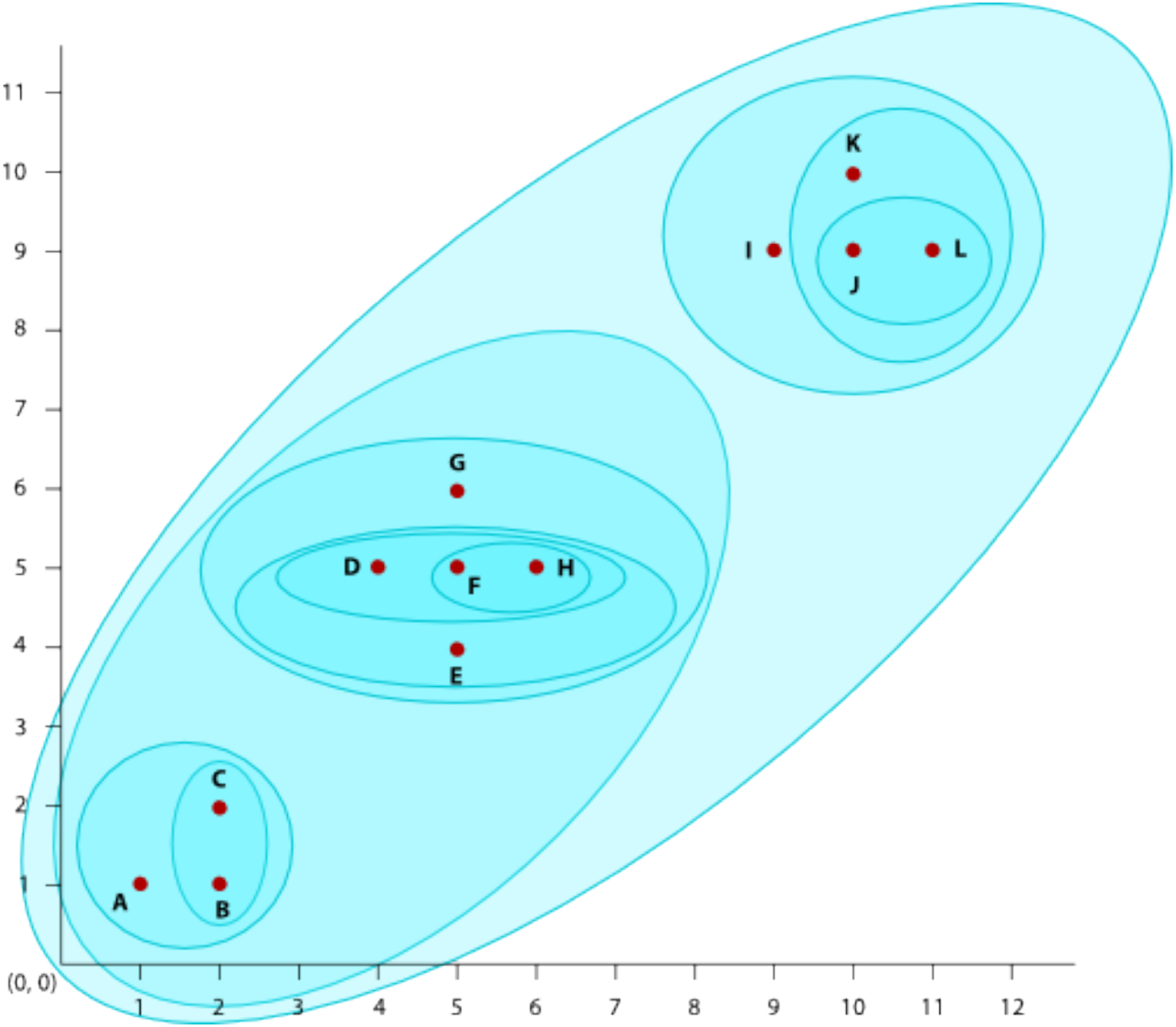
{a},{b}{c},{d}



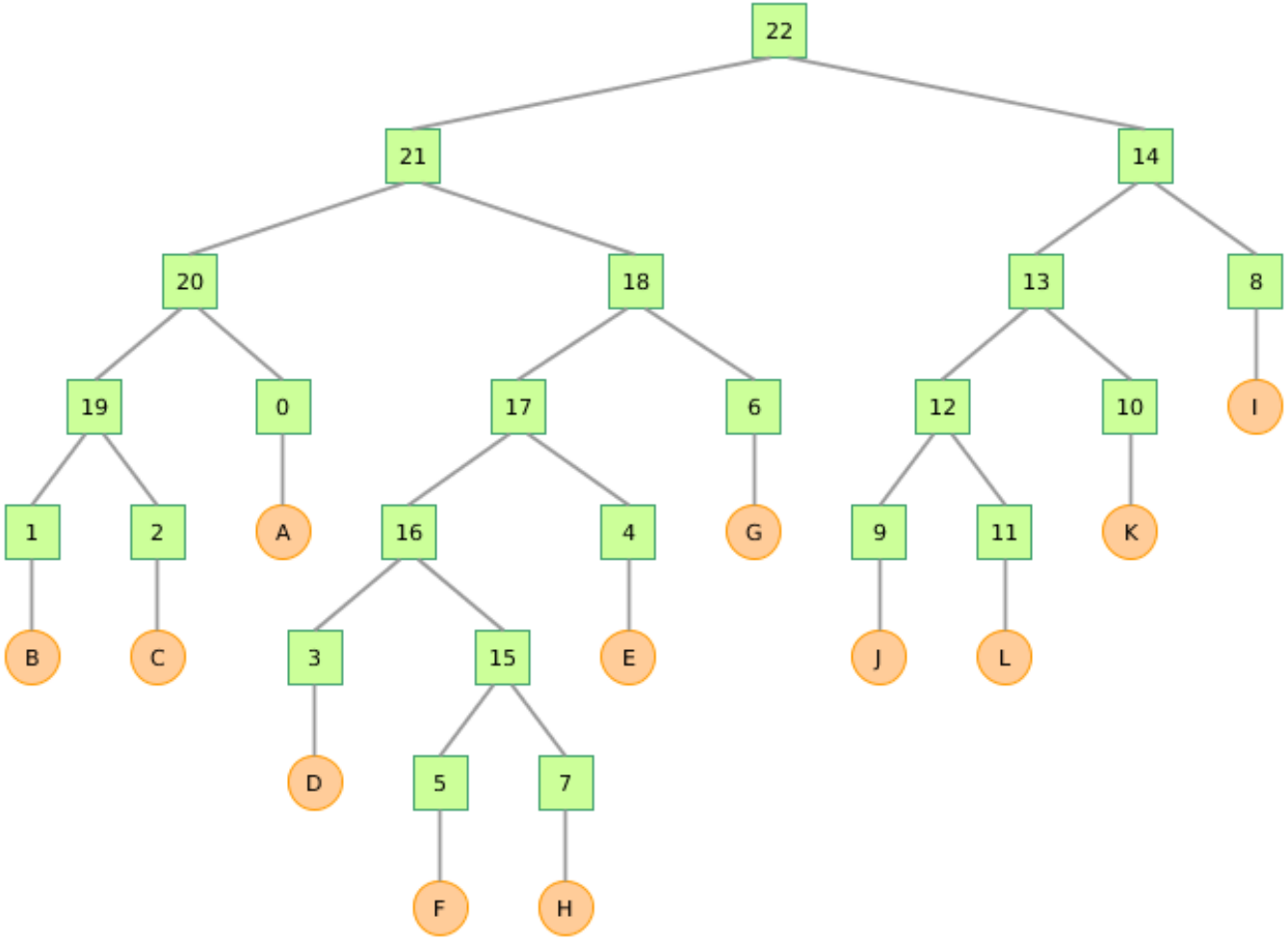
Merge Direction



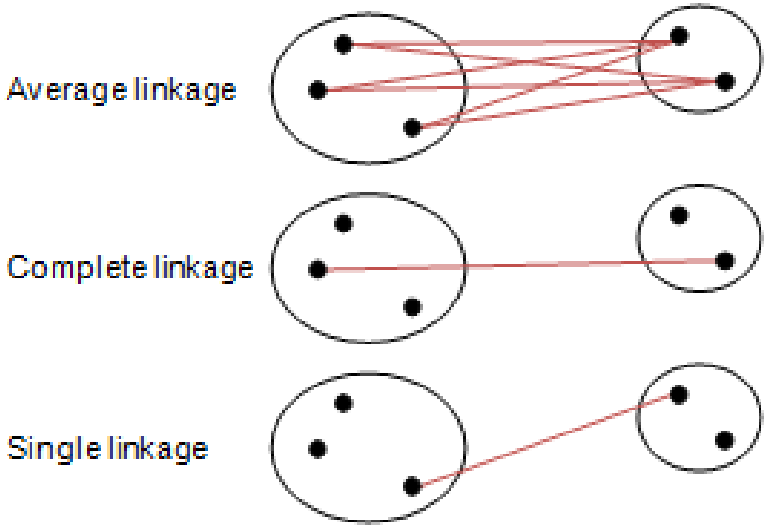
Example



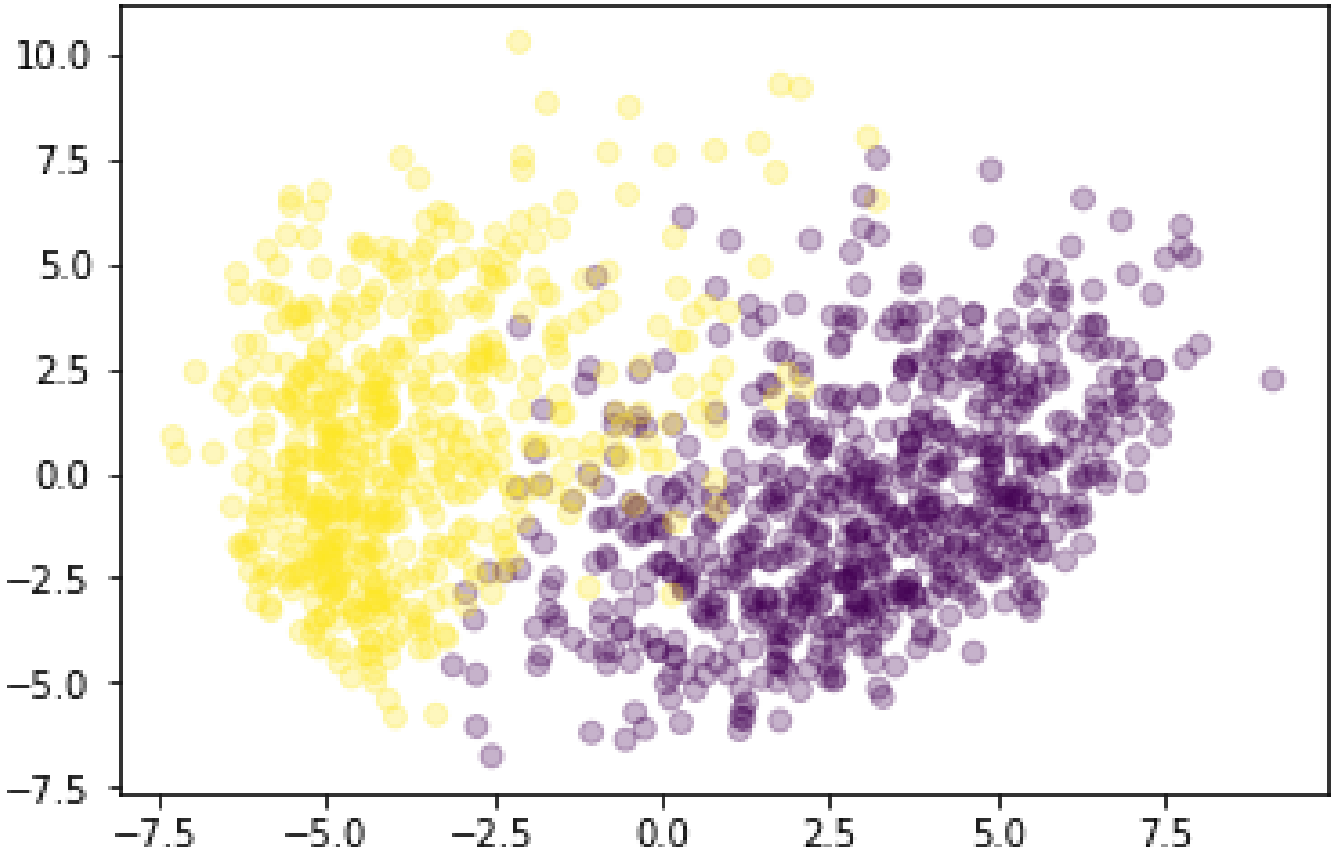
Example



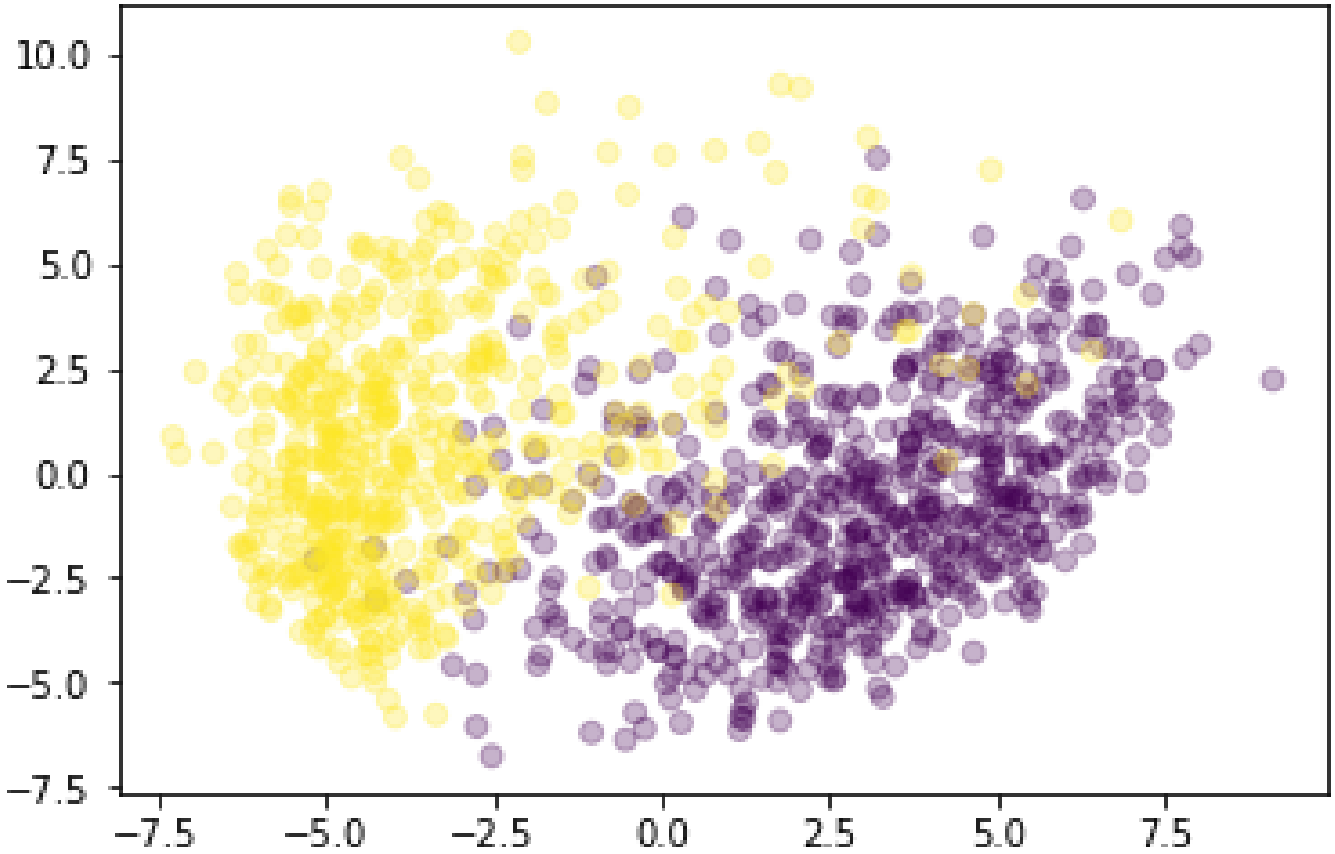
Linkage



Zip Example: Cluster

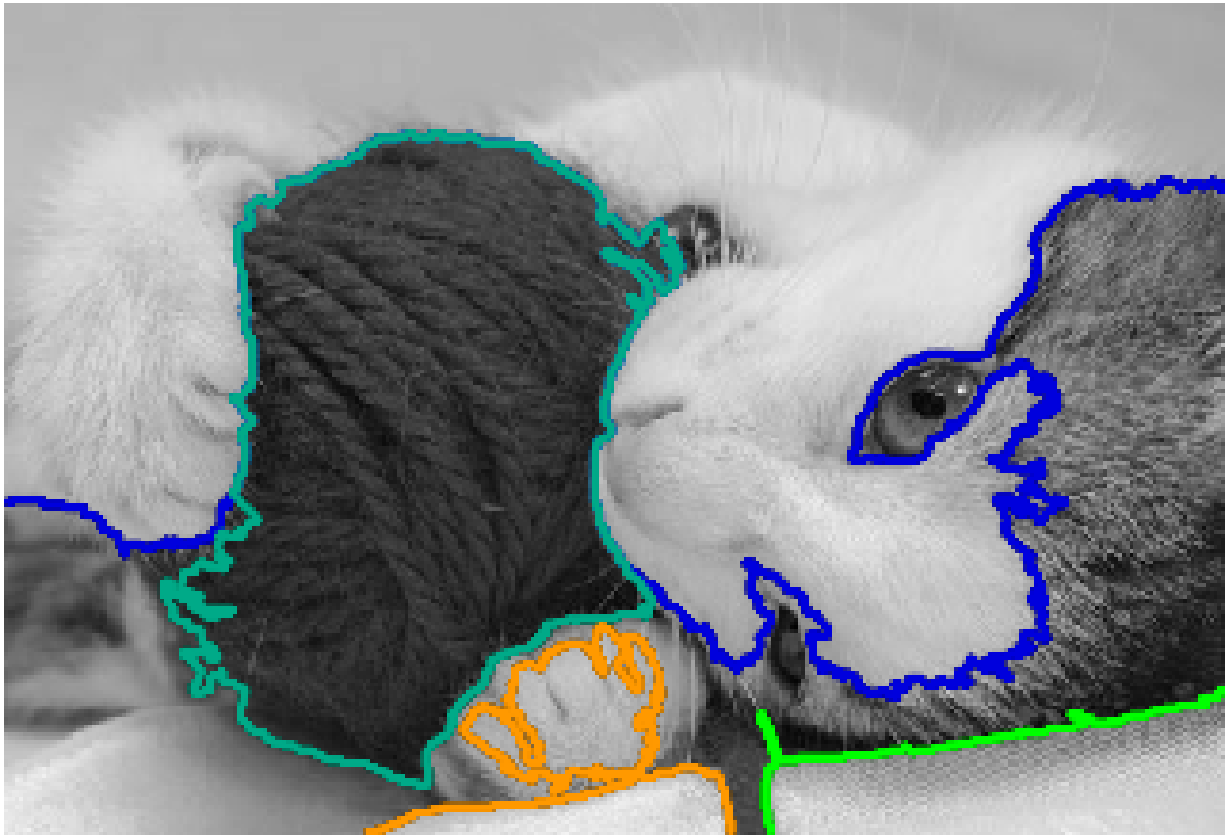


Zip Example: Classify



Cat Example

Cat Example



Cat Example

```
import numpy as np
import matplotlib.pylab as plt
%matplotlib inline
np.random.seed(0)
n = 100
X = np.vstack((np.random.multivariate_normal([0,0], [[1, 0], [0, 1]], n),
               np.random.multivariate_normal([3, 3], [[1, 0], [0, 1]] , n)))
plt.scatter(X[:,0], X[:,1]);
```

```
import scipy.cluster.hierarchy as AHC
X_dist = AHC.distance.pdist(X) #make the pairwise distance
X_linkage = AHC.linkage(X_dist, method = "average")
den = AHC.dendrogram(X_linkage, no_labels = True)
```

```
from sklearn import cluster
X_kmeans = cluster.KMeans(n_clusters = 2)
X_kmeans.fit(X)
X_labels = X_kmeans.labels_
X_centers = X_kmeans.cluster_centers_
```

```
index=X_labels == 0
plt.scatter(X[index, 0], X[index,1], color = "r")
plt.scatter(X[~index,0], X[~index,1], color = "b")
plt.scatter(X_centers[:,0], X_centers[:,1], color = "y",
            facecolor="y", marker="+", s = 700)
```