Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering
IDP Institut für Datenanalyse
und Prozessdesign

**HS 2016**

**Statistisches Data Mining (StDM)**

# Woche 5

## Aufgabe 1 Lab

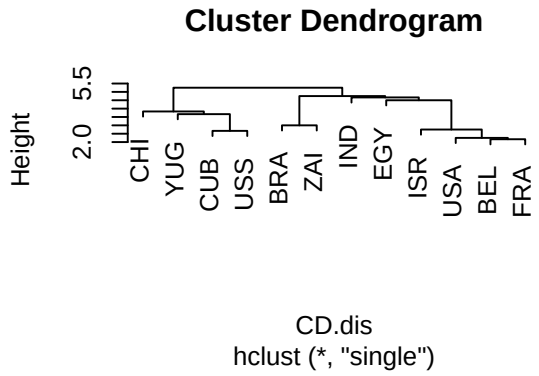
Read and do the excersises of chapter 10.5.2 in ILSR

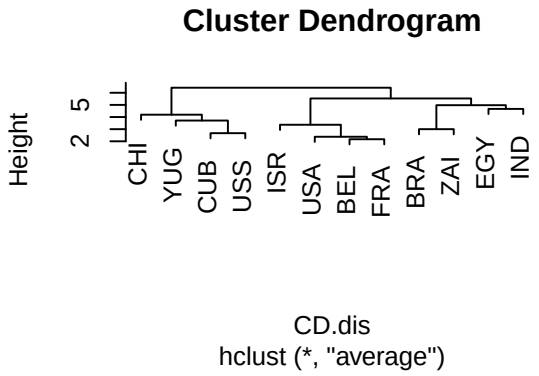
## Aufgabe 2 Clustering und MDS

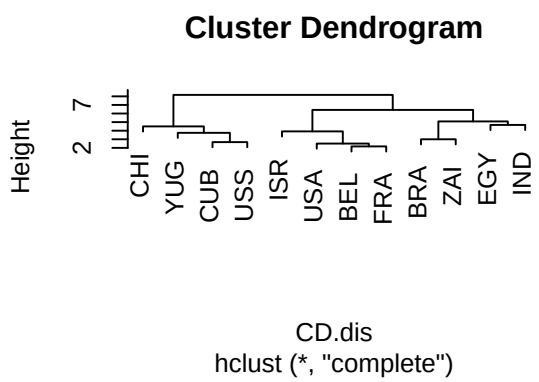
Die Unahnlichkeitsmatrix CD.dis im File CountriesDis.RDA (mit load() laden) enthält Unähnlichkeiten zwischen Ländern. Die Unähnlichkeiten stammen aus einer Studie, in der Studierende aufgefordert waren, paarweise die Unaehnlichkeit zwischen den 12 Ländern Belgien (BEL), Brasilien (BRA), Chile (CHI), Kuba (CUB), Ägypten (EGY), Frankreich (FRA), Indien (IND), Israel (ISR), Vereinigte Staaten (USA), Sowjetunion (USS), Jugoslawien (YUG) und Zaire (ZAI) anzugeben. In der Unähnlichkeitsmatrix sind die durchschnittlichen Unähnlichkeitsbewertung der Studierenden festgehalten.

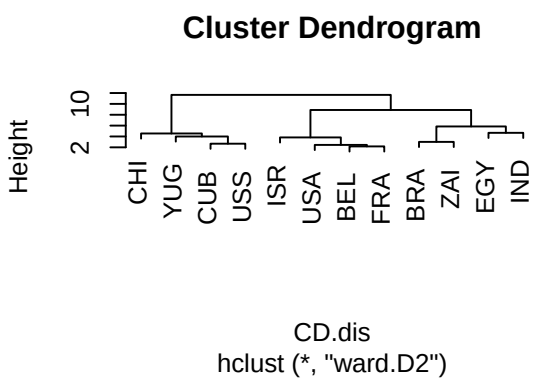
a) Führen Sie mit dieser Unähnlichkeitsmatrix hierarchische Cluster-Analysen durch, verwenden Sie verschiedene Linkage-Methoden. Vergleichen Sie dabei die Resultate der Cluster-Methoden single, complete, average und “ward” bezüglich der Gruppenbildung

```
filename = file.path(baseDir,'CountriesDis.RDA')
load(filename)
c.single = hclust(CD.dis, method="single")
c.average = hclust(CD.dis, method="average")
c.complete = hclust(CD.dis, method="complete")
c.ward = hclust(CD.dis, method="ward.D2")
par(mfrow=c(2,2))
plot(c.single)
plot(c.average)
plot(c.complete)
plot(c.ward)
```



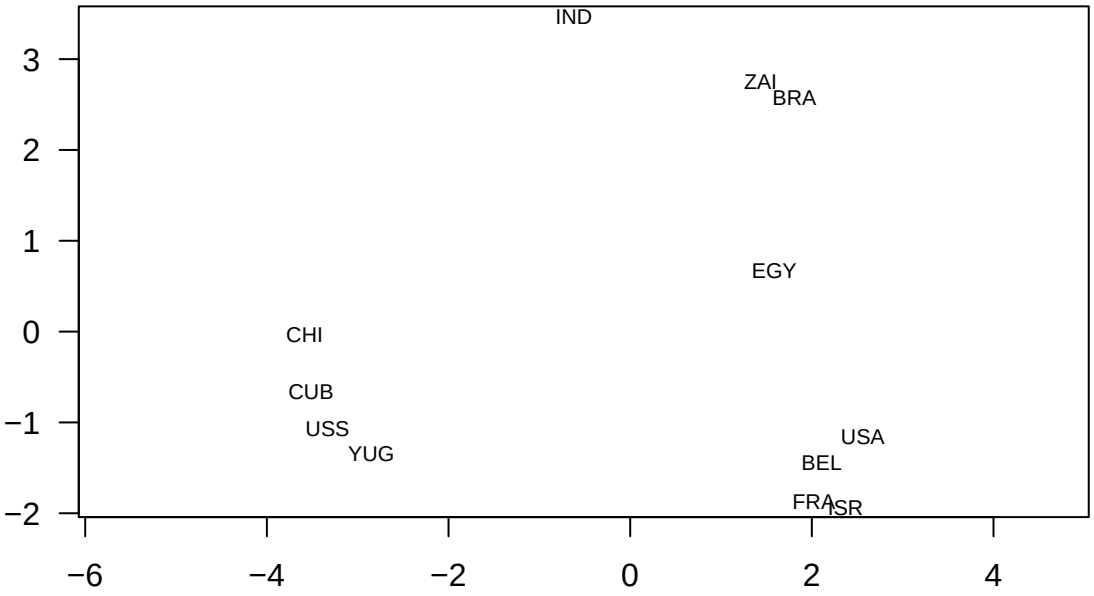

```
  par(mfrow=c(1,1))
  # Clusterbildung bei
# * Single L: Die Luecke ist bei 3.9 bis 4.4. Schneidet man bei 4, so erhalten
#   wir drei Cluster und 2 Mavericks (IND, EGY). Ein Cluster besteht aus CHI,
#   Yug, Cub, USS (kommunistische Laender), ein zweiter aus BRA, ZAI und ein
#   dritter aus ISR, USA, BEL, FRA (westliche Laender)

# * Average L: Hier ist nicht so klar, ob man bei 6 oder bei 4.5 schneiden soll.
#   Schneidet man bei 4.5 gibt es dieselben Cluster wie bei Single L.

# * Complete L: Hier schneidet man am besten bei 5.5, weil die Luecke schon
#   verhaeltnismaessig gross ist. Das fuehrt zu drei Clustern: Kommunistische
#   Laender, westliche Laender, und die restlichen Laender

# * Ward: Ganz klar bei 8 schneiden, sodass es drei Cluster gibt. Die Cluster
#   sind identisch mit der Complete L Methode

# Die Cluster-Bildung durch die vier Methoden ergibt ein fast einheitliches Bild.
# Bei allen Methoden sehen wir die zwei Cluster "kommunistische Laender" und
# "westliche Laender". Ob die restlichen L??nder einen Cluster bilden oder in zwei
# zerfallen, haengt jedoch von der verwendeten Methode ab.
```

b) Führen Sie eine ordinale multidimensionale Skalierung mit zwei Komponenten durch und tragen Sie die Länder in eine 2D-Darstellung ein. Beschreiben Sie die wesentlichen Eigenheiten der Daten, wie sie in dieser Darstellung ersichtlich sind. Wie weit finden Sie darin die Resultate der hierarchischen Cluster-Analyse wieder? Scheiden Sie dazu den Baum aus a) in eine geeignete Anzahl von Gruppen und färben das MDS Ergebniss ein.

```r
  library("MASS")
  CD.kas <- isoMDS(CD.dis, k=2)
```

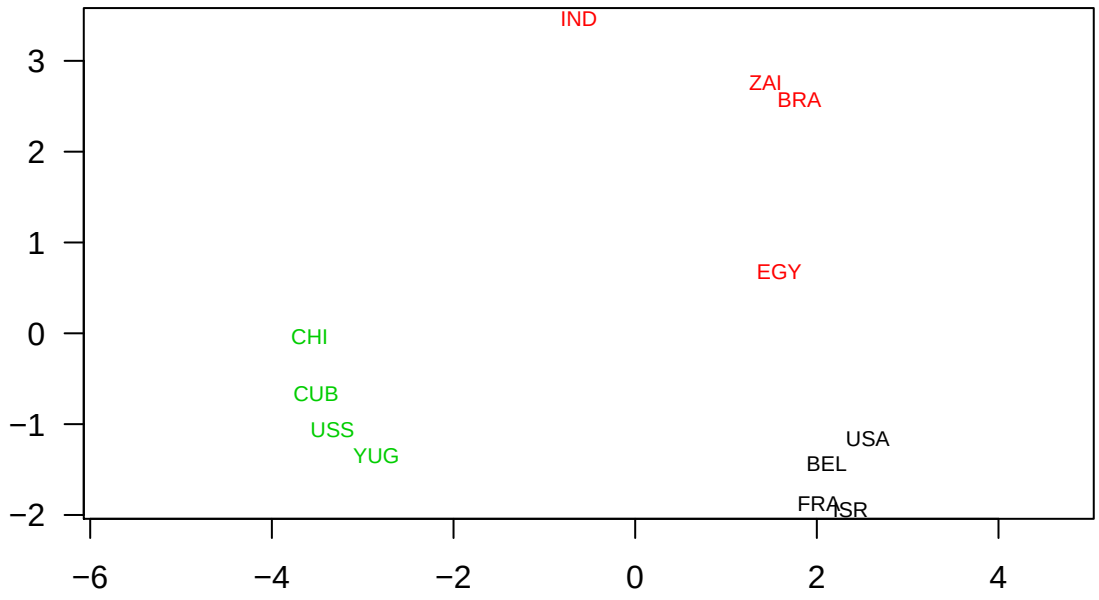
```
## initial  value 11.171833
## iter   5 value 8.475005
## iter   5 value 8.471784
## iter   5 value 8.470026
## final  value 8.470026
## converged
```

```r
par(mfrow=c(1,1))
eqscplot(CD.kas$points, type="n", xlab="", ylab="", cex=0.7, las=1,
         sub="Ordinale MDS nach Kruskal and Shepard")
text(CD.kas$points, labels=labels(CD.dis), cex=0.7, las=1)
```




Ordinale MDS nach Kruskal and Shepard

```r
c.ward = hclust(CD.dis, method="ward.D2")
cluster.k3 = cutree(c.ward, k=3)

#erst plot ohne Punktsymbole erzeugen
eqscplot(CD.kas$points, type="n", xlab="", ylab="",
         cex=0.7, las=1,
         sub="Ordinale MDS nach Kruskal and Shepard")
# jetzt in leeren Plot an den Positionen der Punkte die Namen
# l?nder einf?gen, wobei die Farbe der Clusterzugeh?rigkeit entspricht
text(CD.kas$points, labels=labels(CD.dis), cex=0.7, las=1,
     col=as.numeric(cluster.k3))
```




Ordinale MDS nach Kruskal and Shepard

```
# Die Punkte liegen auf zwei "Bananen". In der linken sind die vier Punkte sehr
# regelmaessig verteilt. Es handelt sich um die kommunistische Laender.
# In der rechten Bananen hat es weiter Unterstrukturen. Unten liegen die
# westlichen Laender, dann folgt gegen oben EGY, dann das Duo BRA und ZAI und am
# oberen Ende liegt IND.
# Die beiden Cluster mit den kommunistischen und den westlichen Laender findet
# man wieder. Die restlichen Laendern sind nochmal anders gegliedert.
```

## Aufgabe 3 Clustering und MDS

In dieser Aufgabe untersuchen wir nochmals die Daten zu den letzten 20 eidgenössischen Volksabstimmungen von 1998 und 1999.

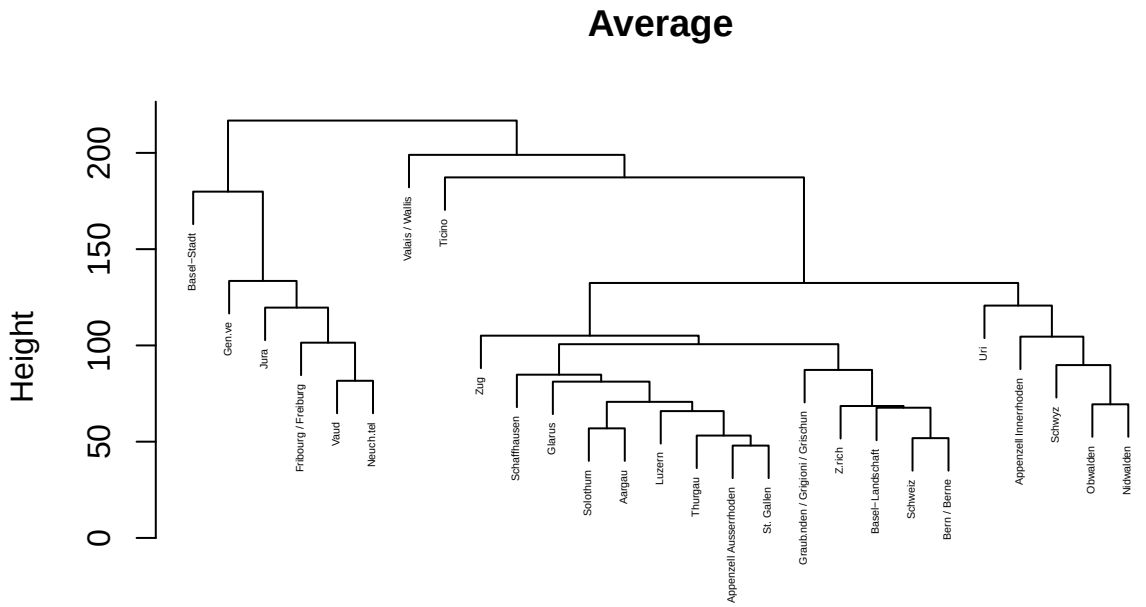
a) Führen Sie eine hierarchische Cluster-Analyse durch unter Verwendung von euklidischen Distanzen zwischen den Kantonen. Vergleichen Sie die Resultate der Cluster-Methoden Single Linkage, Complete Linkage und Average Linkage bezüglich der Gruppenbildung. Was ist jeweils eine optimale Anzahl Cluster? Wie verhält sich in den Resultaten jeweils die Schweiz (CH)? Tipps: Die Daten können Sie wie folgt einlesen:

```
X.t <- read.table(file.path(baseDir, "../PCA/px-x-1703030000_100.csv"),
                  header=T, sep=";", skip=1, stringsAsFactors = FALSE)
X <- reshape(X.t, idvar = "Kanton", timevar = "Datum.und.Vorlage", direction = "wide")
```

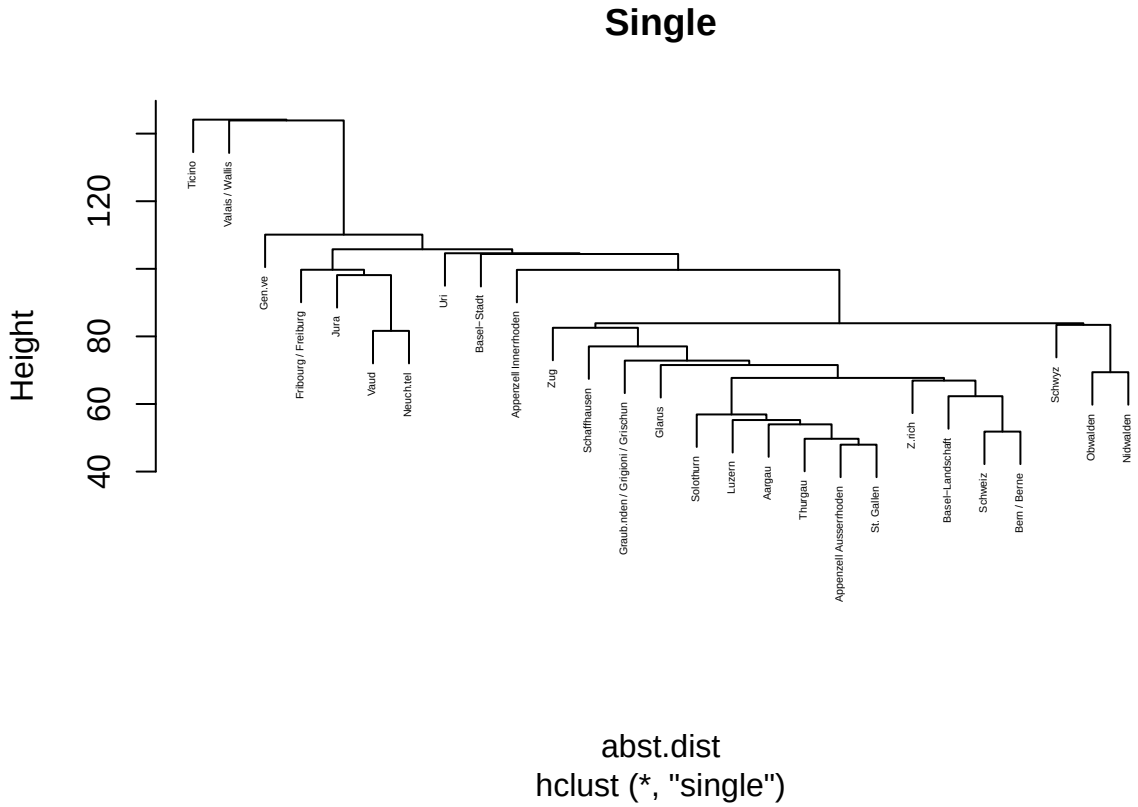
Zur besseren Visualisierung können Sie den Parameter `cex=0.3` setzten.

```
X.t <- read.table(file.path(baseDir, "../PCA/px-x-1703030000_100.csv"), header=T, sep=";", skip=1,
X <- reshape(X.t, idvar = "Kanton", timevar = "Datum.und.Vorlage", direction = "wide")

names = X$Kanton
X = data.matrix(X[,2:ncol(X)])
row.names(X) = names
```

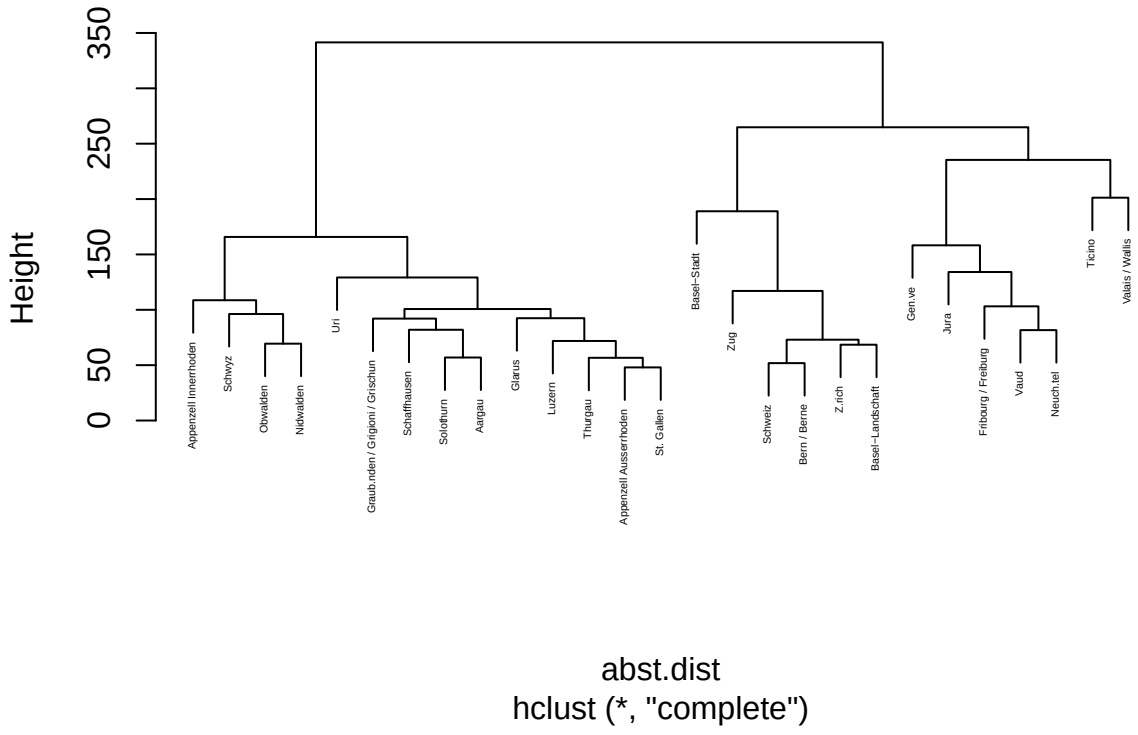
```r
abst.dist <- dist(X)

par(mfrow=c(1,1))
cex.val = 0.3
abst.av <- hclust(abst.dist, method="average")
plot(abst.av, labels=names, main="Average", cex=cex.val)
```



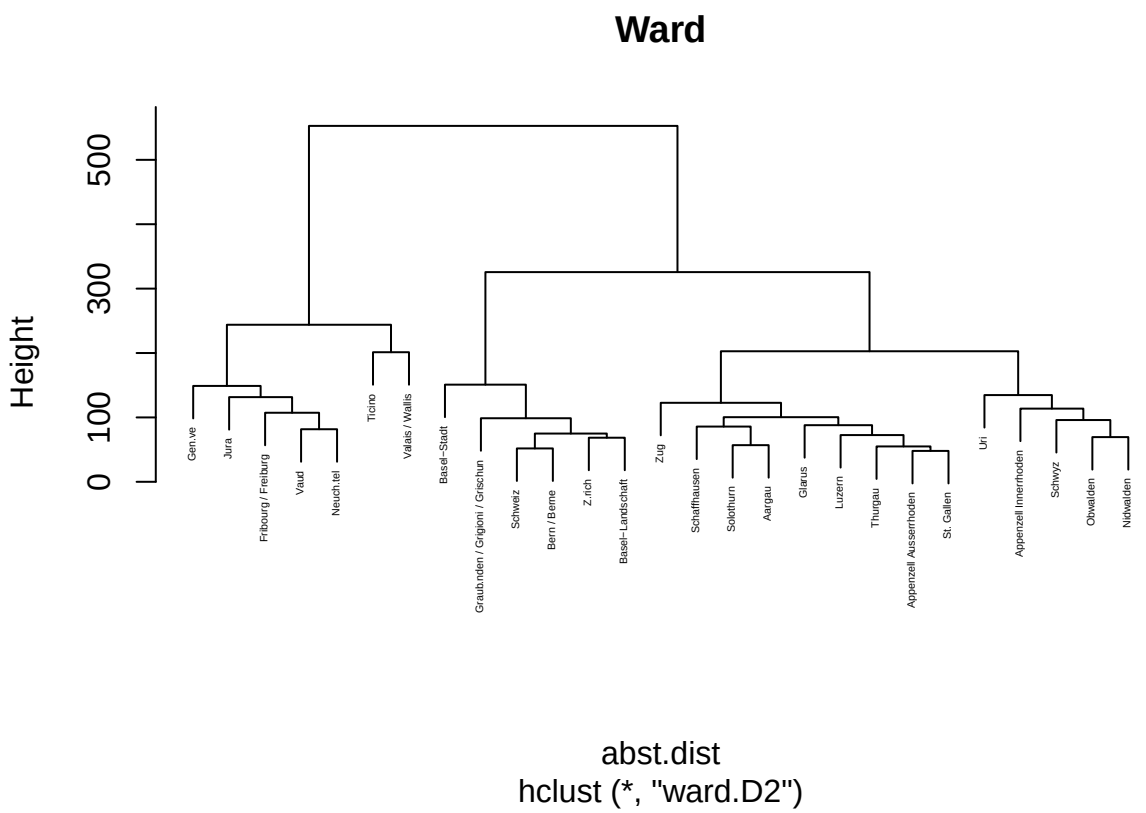

```r
abst.con <- hclust(abst.dist, method="single")
plot(abst.con, labels=names, main="Single", cex=cex.val)
```




```
abst.com <- hclust(abst.dist, method="complete")
plot(abst.com, labels=names, main="Complete", cex=cex.val)
```



```r
abst.W <- hclust(abst.dist, method="ward.D2")
plot(abst.W, labels=names, main="Ward", cex=cex.val)
```




```r
# Die Methoden Complete und Ward liefern die "schoensten" Ergebnisse: Vier klare
# Clusters bei Complete Linkage und drei Clusters bei der Methode von Ward.
# Bei den Methoden Average und Single ist die Anzahl Gruppen unklar.
```

b) Führen Sie eine Cluster-Analyse mit K-means (K=3) durch.

```r
abst.km0 <- kmeans(abst.dist, centers = 3)
abst.km0$cluster # Clu
```

c) Wählen Sie die optimale Anzahl K-means-Cluster. Um zu sehen, wie stark das Ergebniss vom Startwert abhängt führen sie 5 Verschiedene Durchläufe aus. Verändern Sie auch den Parameter `nstart`, was fällt auf.

```r
getRes = function(){
  x.res <- rep(NA,16)
  # nun wird f?r jede Abstimmung die Varianz der Ergebnisse in den
  # verschiedene Kantonen berechnet - statt einer for-Schleife wird
  # hier mit apply gearbeitet
  for(i in 2:16){
    abst.km <- kmeans(abst.dist, centers = i, nstart=10)
    x.res[i] <- sum(abst.km$withinss)  # bestimme die summe der Varianzen ?ber alle Cluster
  }
  return (x.res)
}
plot(2:16, getRes()[2:16], type="b")
for (i in 1:10){
```

```
  lines(2:16, getRes()[2:16], col='gray')
}
# Schwer zu sehen, gehen wir mal von K=4 aus.
# Wählt man nstart gross, so gleichen sich die Kurven an.
```

d) [Optional] Wählen Sie ein geeignetes K, führen Sie ein K-Mean Clustering durch und färben Sie bei einem hierachischen gemäss der Gruppenzugehörigkeit ein. Tipp: http://stackoverflow.com/questions/18802519/label-and-color-leaf-dendrogram-in-r

```
abst.km <- kmeans(abst.dist, centers = 4, nstart=100)
abst.W <- hclust(abst.dist, method="ward.D2")
plot(abst.W, labels=names, main="Ward", cex=cex.val, col=abst.km$cluster)
# TODO make coloring of nodes
```