

11_clusteranalyse

Gruppierung von Stichproben anhand ihrer
Merkmale



Clusteranalyse: Idee und Grundlagen

Ähnliche Dinge haben ähnliche Merkmale...

Gruppenbildung anhand von Merkmalsausprägungen, die sich (deutlich?) von anderen Gruppen unterscheiden

Intuitiv Grundlage archäologischer Arbeit

Mit Ende 60er Jahre (New archaeology) Wunsch,

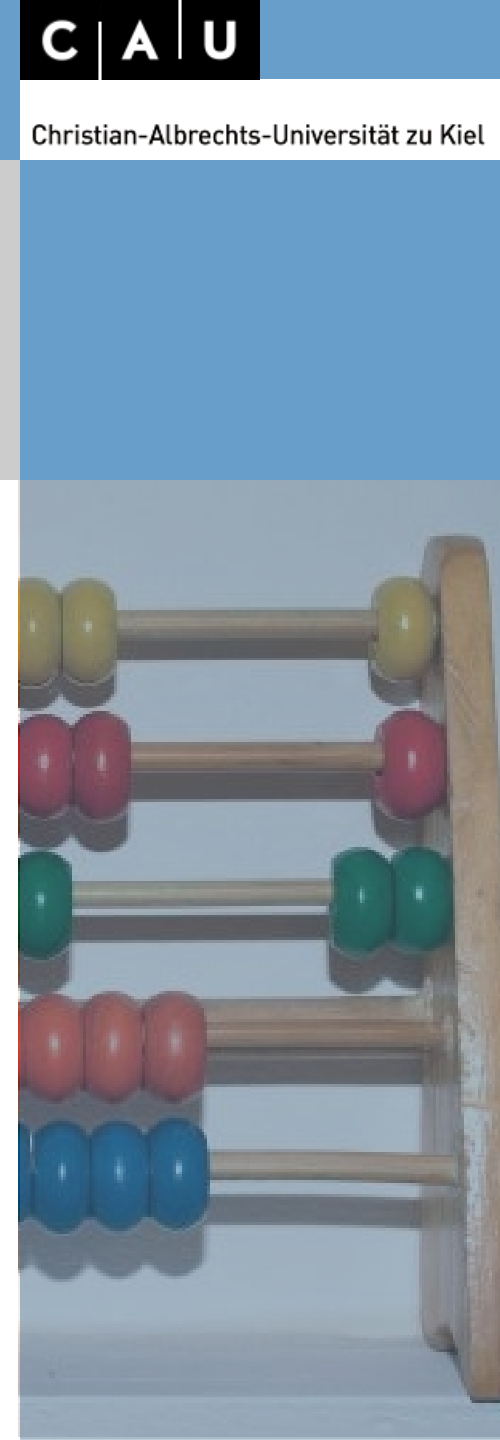
- Kriterien für Gruppenbildung von subjektiven Entscheidungen abzukoppeln
- Verarbeitung von großen, intuitiv unüberblickbaren Datenmengen zu ermöglichen

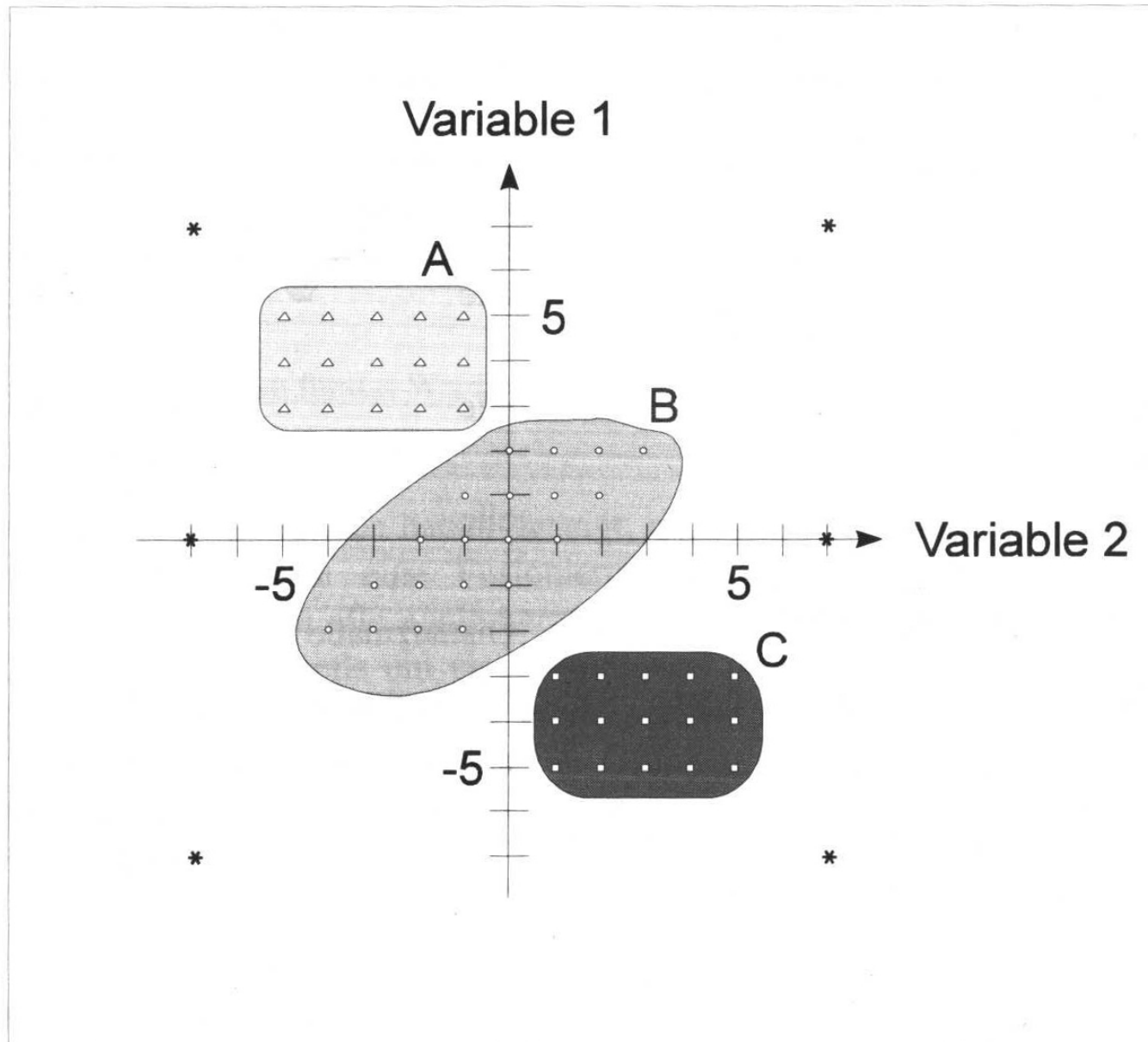
→ multivariate Analysen

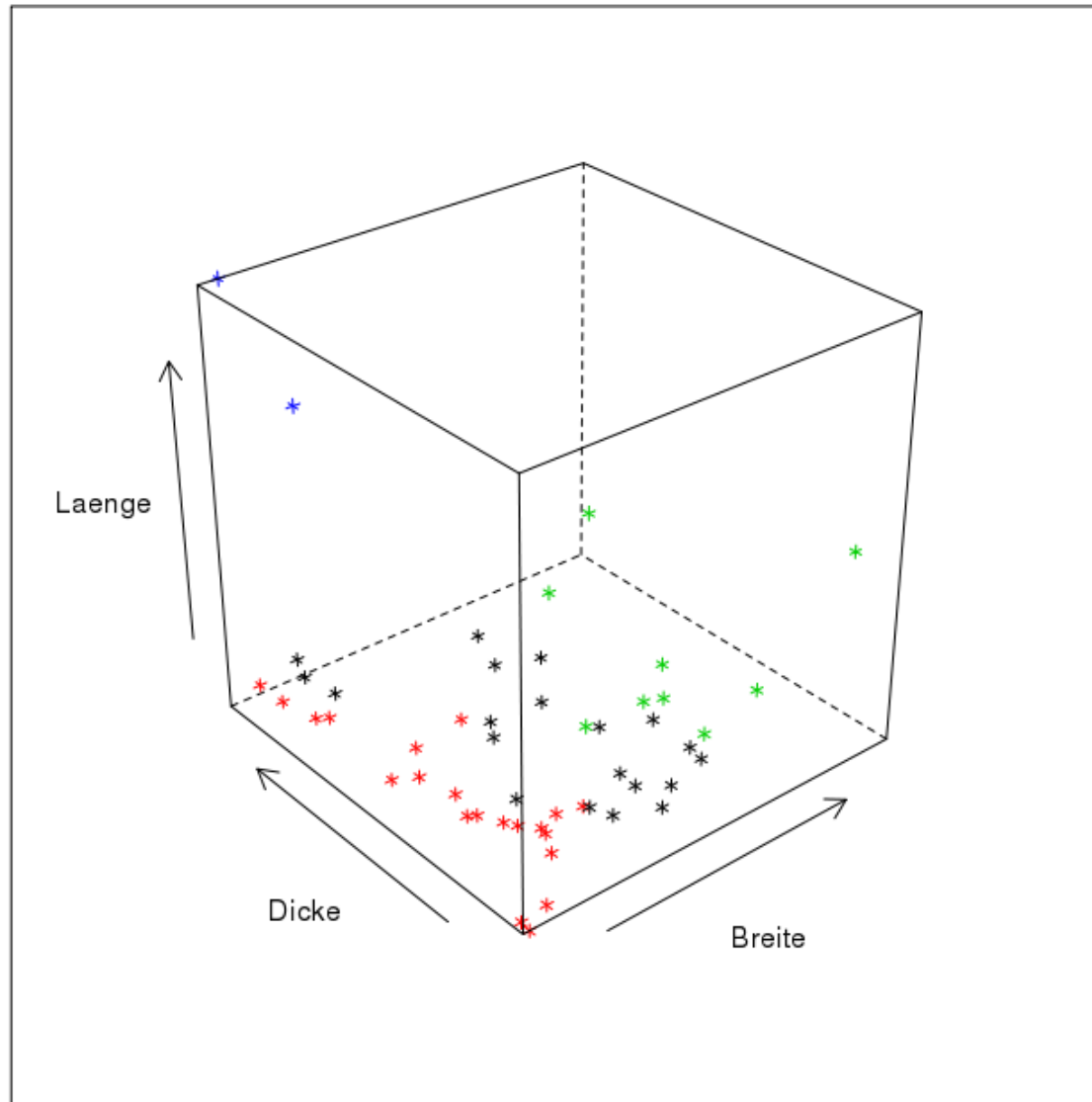
Clusteranalyse

1. Messung eines (wie auch immer gearteten) Abstandes von Daten voneinander
2. Gruppierung von Daten, die sich ähneln, gegenüber Daten, die sich unterscheiden

→ **Klassifikation**







Clusteranalyse: Methoden [1]

Getrennt marschieren, vereint schlagen... oder?

Hierarchisch

Welche Objekte sind sich am ähnlichste?

Welche Objekte sind sich am 2. ähnlichsten?

Welche Objekte sind sich am 3. ähnlichsten?

...

→ **agglomerativ**

Gehe von der kleinsten Einheit aus (einzelne Objekte)

Fasse die beiden ähnlichsten zu einem Objekt zusammen (1. Cluster)

Fasse die beiden ähnlichsten [Cluster|Objekte] zusammen

...

→ **unterteilend**

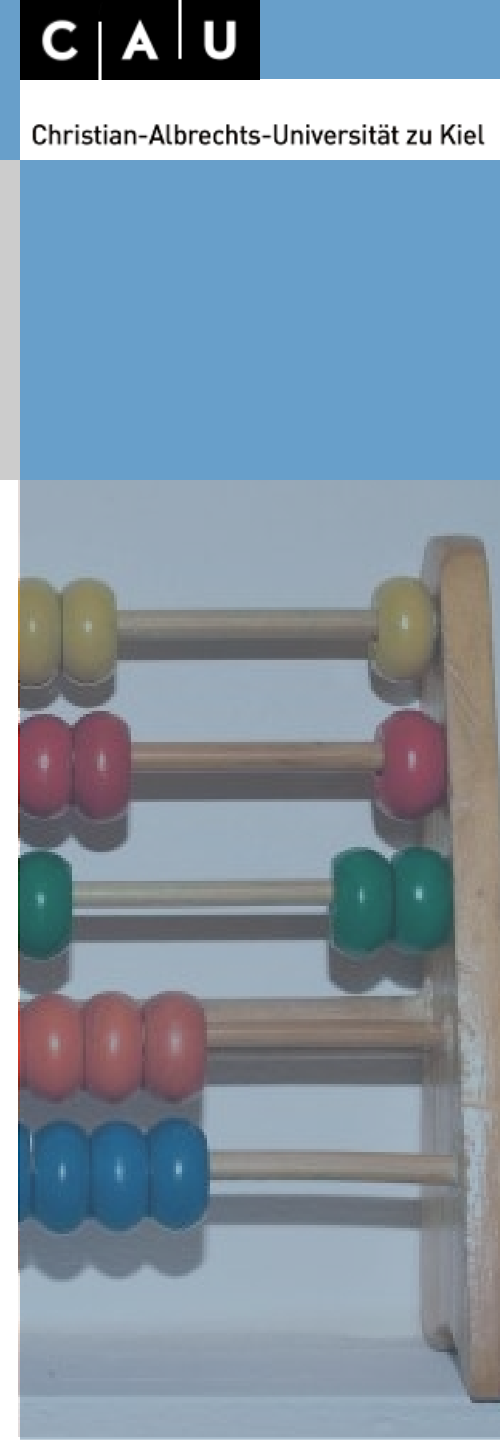
Fange mit der größtmöglichen Einheit an (alle Objekte als 1 Cluster)

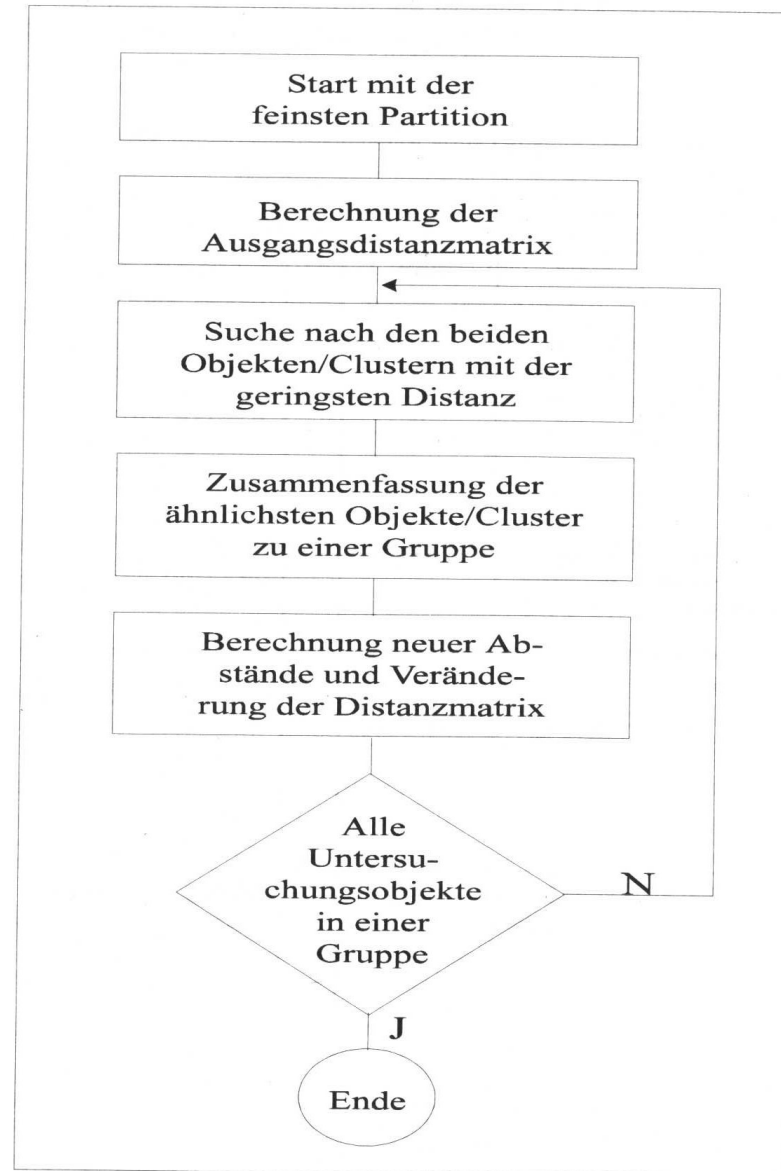
Teile diesen in zwei möglichst unähnliche Gruppen

Teile eine der Gruppen in zwei möglichst unähnliche Gruppen

...

Hierarchisches Clustern, z.B. nach Ward-Methode





Clusteranalyse: Methoden [2]

Teile und herrsche... oder?

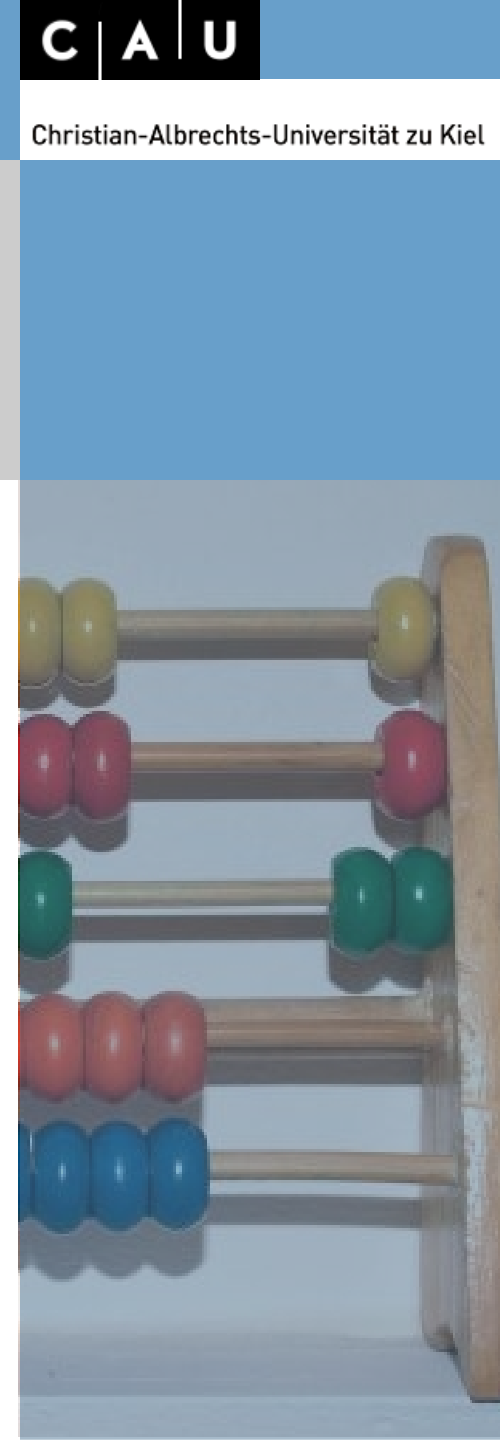
Partitionierend

Wie lassen sich die Daten am sinnvollsten in n Gruppen unterteilen?

Möglicher Ablauf:

1. Wähle n Clusterzentren zufällig aus.
2. Fasse Daten zusammen, die diesen Clusterzentren am ähnlichsten sind
3. Berechne die Clusterzentren ggf. Neu
4. Ändert sich was?
Wenn ja, wieder zu 2.
Sonst: fertig!

kmeans-Clustering



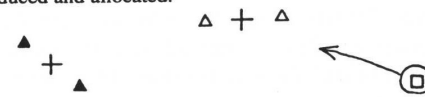
1. Two individuals are selected as starting points for the two clusters; a third individual is introduced and allocated to its nearest cluster:



2. The position of the centre of cluster 2 is recalculated; another case is introduced and allocated:



3. Position of centre of cluster 1 is recalculated; another case is introduced and allocated:



4. Position of centre of cluster 2 is recalculated; another case is introduced and allocated:



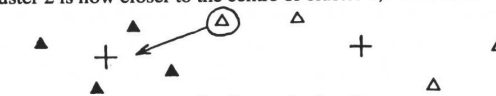
5. Position of centre of cluster 1 is recalculated; another case is introduced and allocated:



6. Position of centre of cluster 2 is recalculated; another case is introduced and allocated:



7. Position of centre of cluster 1 is recalculated; left-most individual of cluster 2 is now closer to the centre of cluster 1, so is allocated to it:



8. Centres of both clusters finally recalculated;

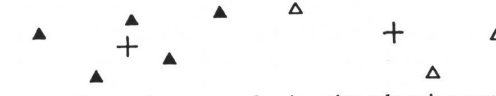


FIGURE 11.7. Successive stages of an iterative relocation partitioning procedure for two clusters.

Clusteranalyse: Methoden [3]

Vorteile-Nachteile

Hierarchisch

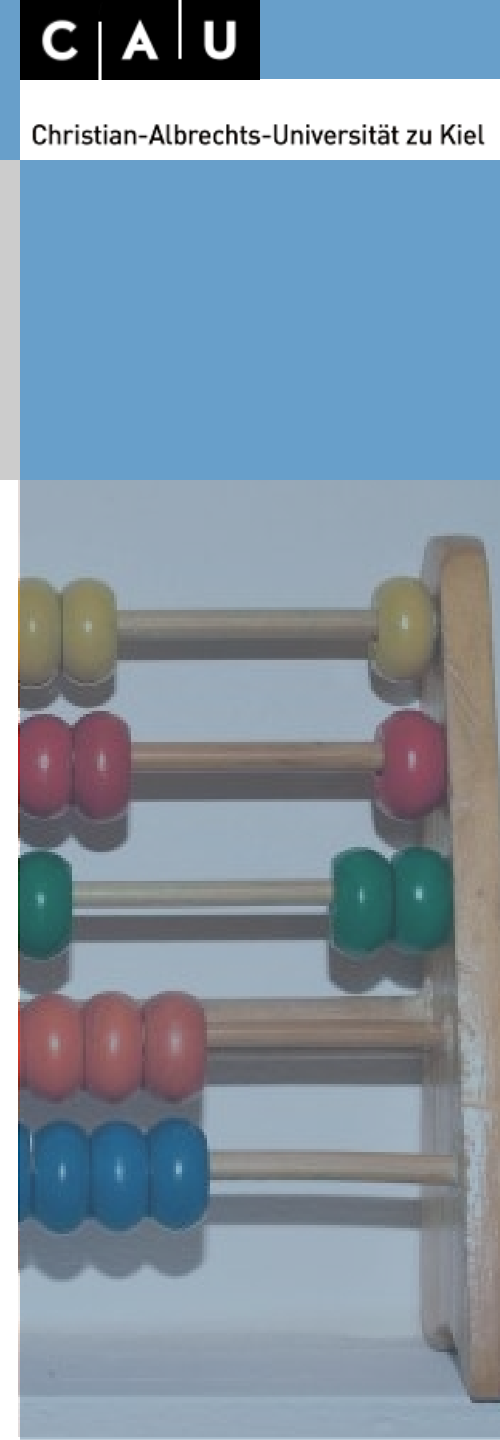
Vorteil: Es wird keine Anzahl von Clustern vorgegeben, Hierarchien von Clustern sind beobachtbar (Darstellung in einem Dendrogramm)

Nachteil: Einmal gefundene Lösung kann nicht wieder aufgelöst werden, auch wenn der Cluster in einem späteren Schritt nicht mehr optimal ist.

Partitionierend

Vorteil: Cluster sind im nachhinein noch variabel, d.h., wenn sich nach einem Clusterdurchgang eine bessere Lösung findet, kann diese gewählt werden

Nachteil: Es wird eine Clusterzahl vorgegeben.



Distanzberechnungen: Euklidische Distanz (metrische Variablen)

Der Kürzeste Weg zwischen zwei Punkten ist immer noch die Gerade

Je näher zwei Punkte zueinander sind, deren Position in einem Koordinatensystem durch die Werte der jeweiligen Variablen bestimmt wird, umso ähnlicher sind sich die Datensätze.

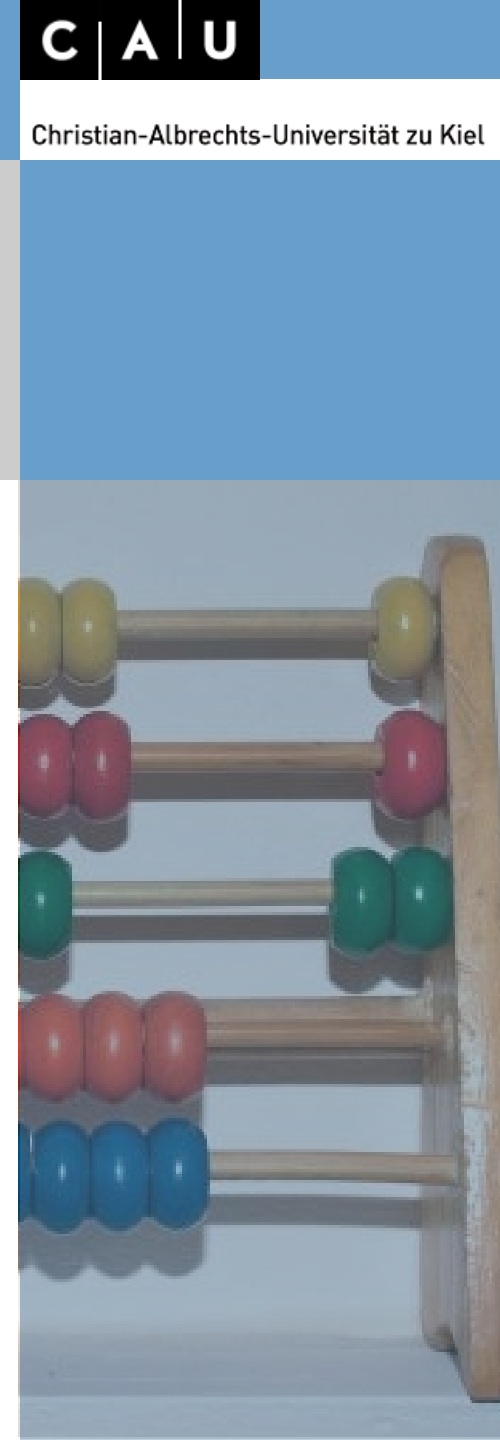
Berechnung der Nähe zueinander:
Satz des Pythagoras...

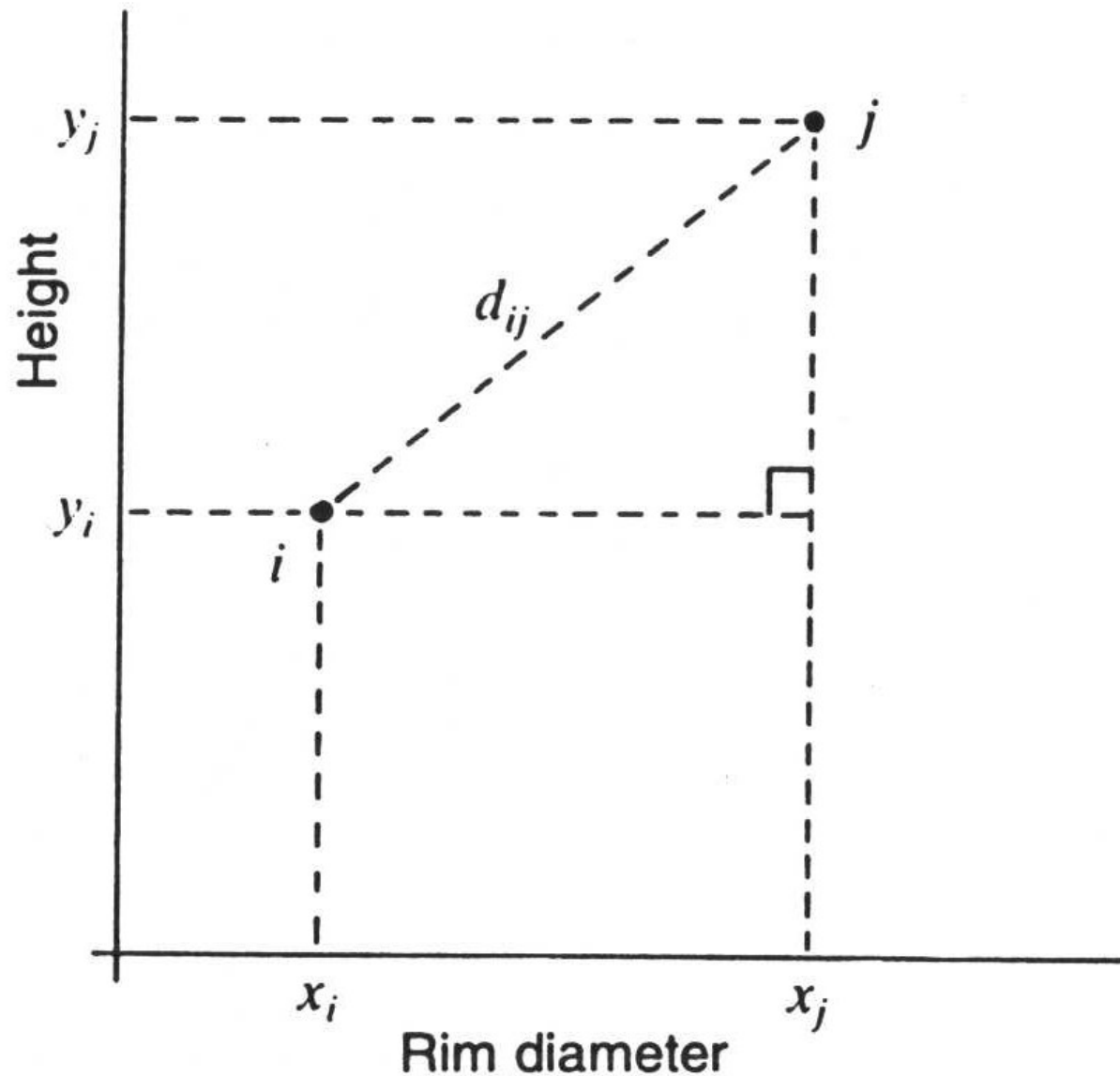
$$a^2=b^2+c^2$$

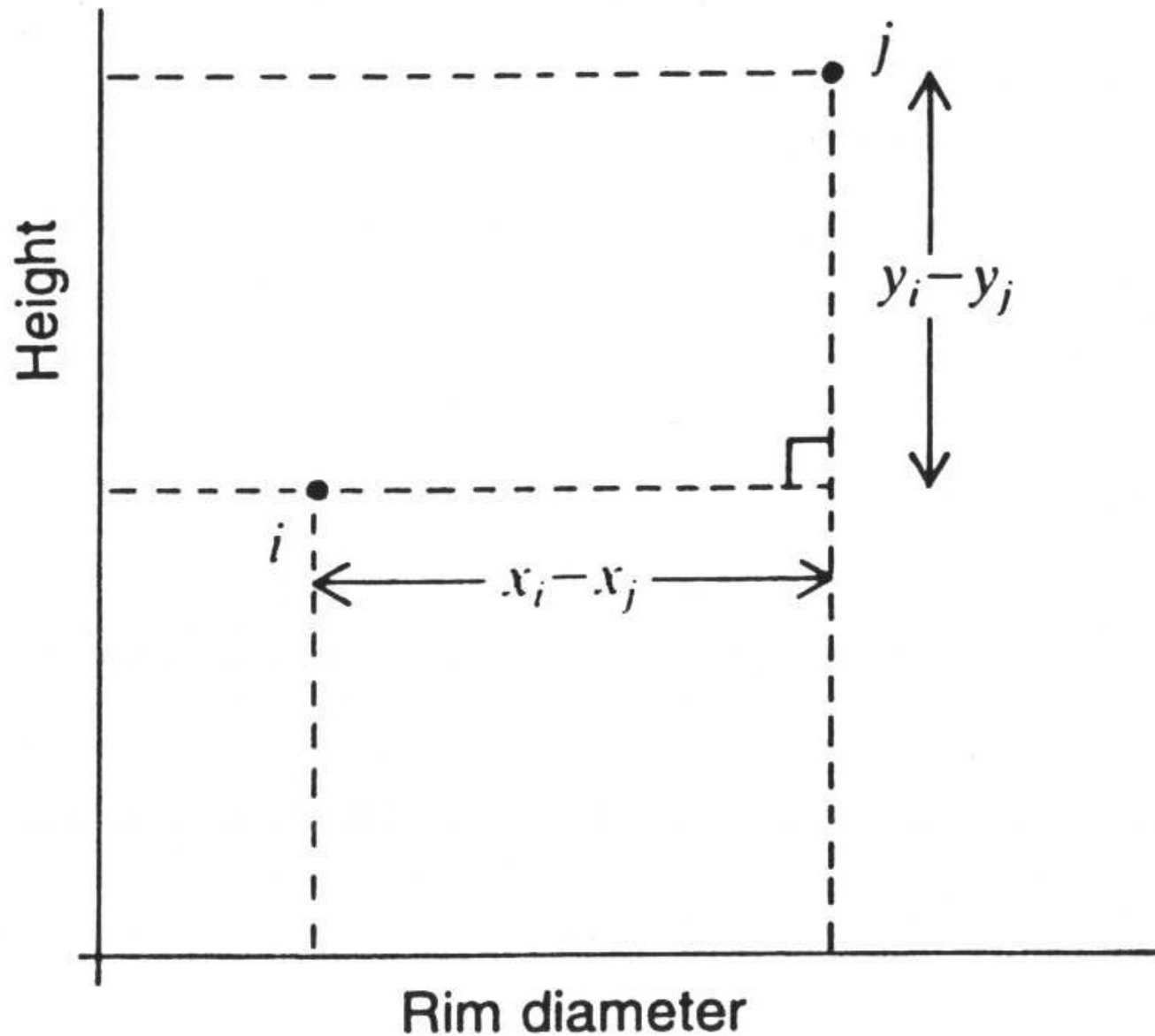
Der Abstand zweier Daten mit den Variablen x,y ist also:

$$d_{ij}=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}$$

Gilt auch für mehr als zwei (drei, viele) Dimensionen







Distanzberechnungen: City-Block Distanz (oder auch Manhattan-Metrik) (metrische Variablen)

Wie der Taxifahrer fährt

Wiedergabe der absoluten Distanz zweier Objekte zueinander

Problem: Wenn die zwei Variablen irgendwie voneinander abhängig sind, ist das entstehende Koordinatensystem nicht rechtwinklig.

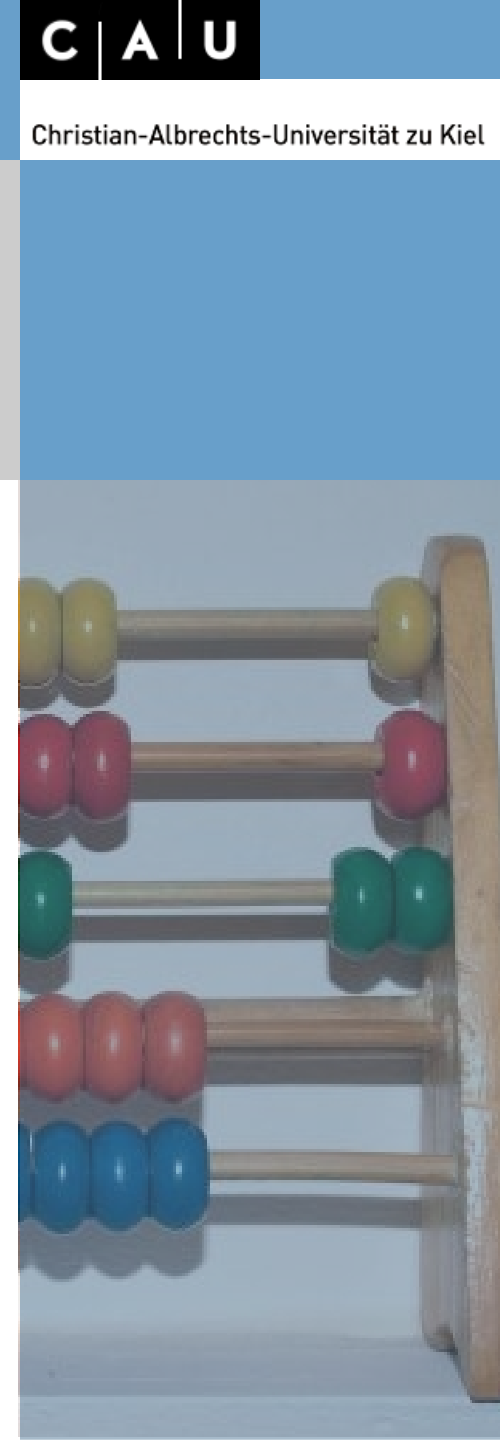
Daher würden mit euklidischer Metrik Entfernungen über- oder unterschätzt.

Lösung: City-Block-Distanz

Der Abstand zweier Daten mit den Variablen x, y ist also:

$$d_{ij} = |x_i - x_j| + |y_i - y_j|$$

Gilt auch für mehr als zwei (drei, viele) Dimensionen



Distanzberechnungen: nichtmetrische Variablen (Präsenz-/Absenzzmatrizen) [1]

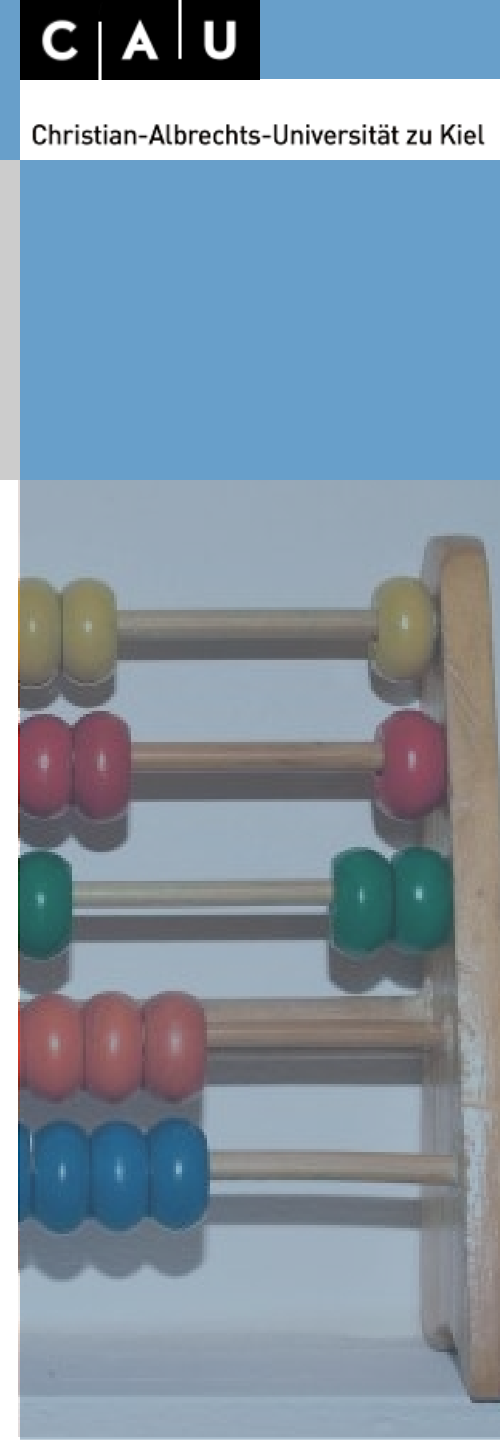
Wenn Abstände nicht mehr berechnet werden können

Bei nominalen oder ordinalen Variablen gibt es keine definierten Abstände mehr zwischen den Werte (ist hoffentlich noch bekannt...)

Daher können diese auch nicht mehr im euklidischen Raum berechnet werden.

Mögliche Lösungen: Berechnung über Ähnlichkeitskoeffizienten aus Kontingenztabellen. Bsp: Grabinventare

Grab 1	Grab 2	
	+	-
+	a	b
-	c	d



Distanzberechnungen: nichtmetrische Variablen (Präsenz-/Absenzmatrizen) [2]

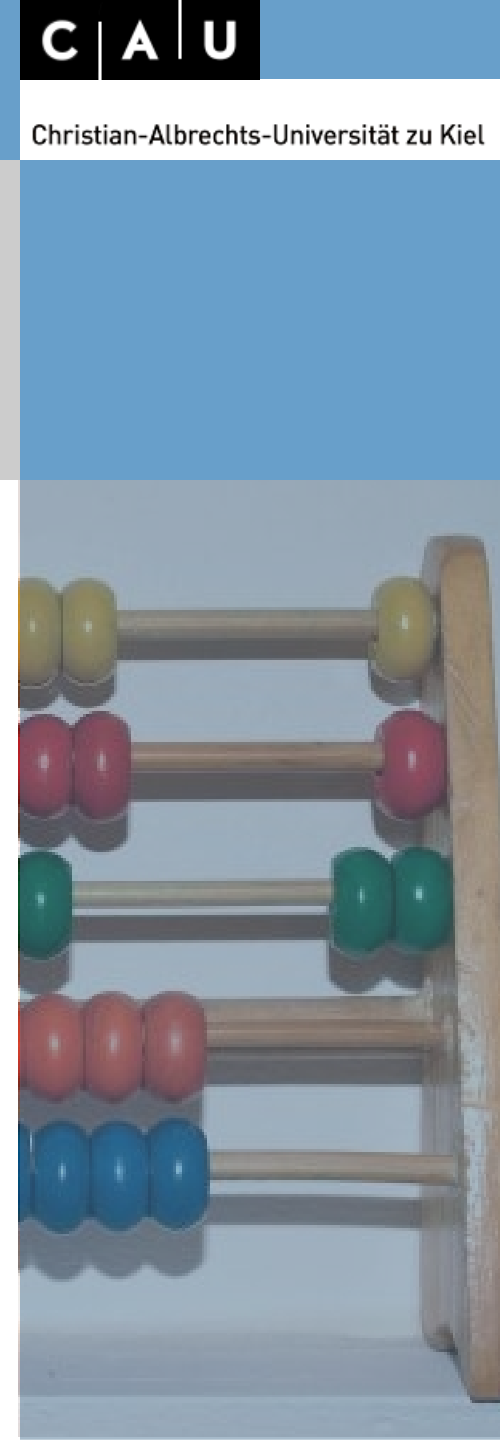
Berechnung der Ähnlichkeiten über gleich/unterschiedliche Merkmale

Geprüft wird, in wievielen Fällen die Gräber übereinstimmen (a,d) und in wievielen Fällen es Unterschiede (b,c) gibt.

	Grab 2	
	+	-
+	a	b
-	c	d

Typen	1	2	3	4	5	6	7	8	9
Grab 1	1	1	0	1	0	0	1	1	1
Grab 2	1	0	0	0	0	0	1	0	1

	Grab 2	
	+	-
+	3	3
-	0	3



Distanzberechnungen: nichtmetrische Variablen (Präsenz-/Absenzmatrizen) [3]

Berechnung der Ähnlichkeiten über gleich/unterschiedliche Merkmale

Verschiedene Möglichkeiten, die Abstände zu berechnen:

$$\text{Tanimoto (Jaccard): } d = \frac{a}{a+b+c}$$

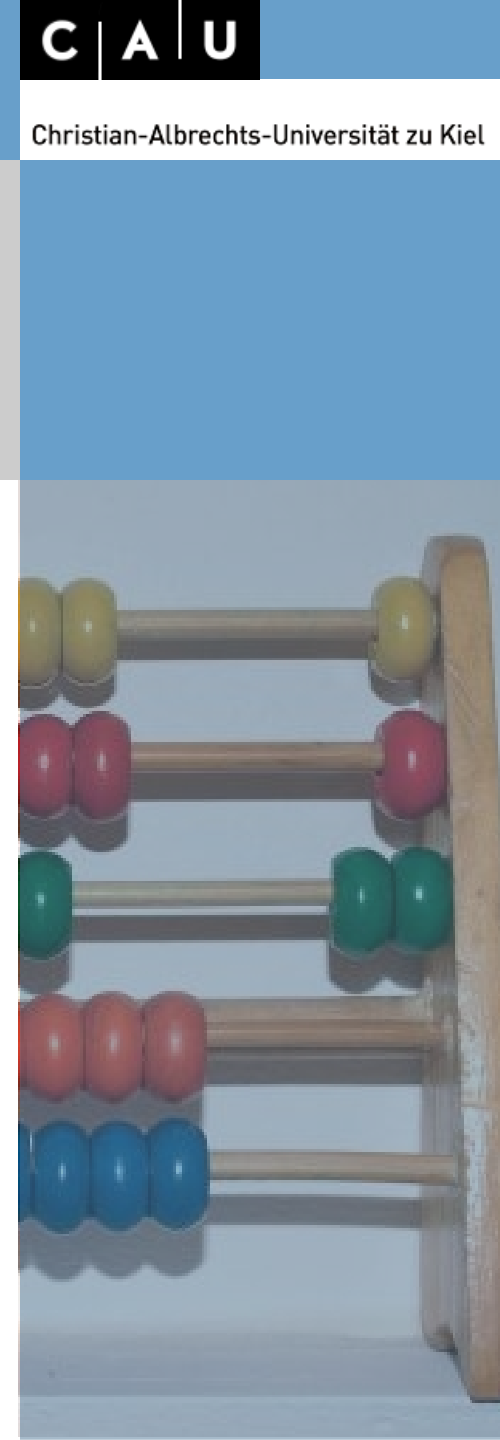
$$\text{Simple Matching: } d = \frac{a+d}{a+b+c+d}$$

$$\text{Russel \& Rao (RR): } d = \frac{a}{a+b+c+d}$$

Beispiel Gräber mit Tanimoto (Jaccard):

$$d = \frac{3}{3+3+0} = \frac{3}{6} = 0.5$$

Grab 1	Grab 2	
	+	-
+	3	3
-	0	3



Distanzberechnungen: nichtmetrische Variablen (Präsenz-/Absenzzmatrizen) [4]

Das ganze in R

```
> leder <- read.csv2("leder_julia.csv")

> dist(leder[,c("Breite", "Laenge", "Dicke")], method="euclid")
...

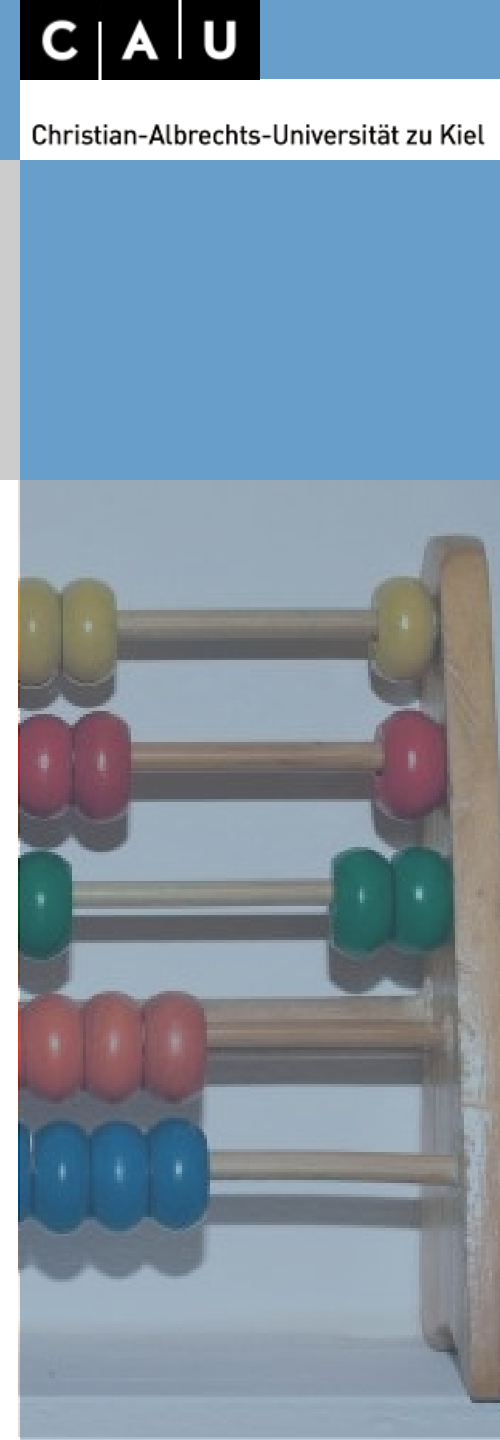
> dist(leder[,c("Breite", "Laenge", "Dicke")], method="manhattan")
...

> graeber <- read.csv2("graeber_pa.csv", row.names=1)

> library(vegan)

> vegdist(graeber, method="jaccard")
```

	grab1	grab2	grab3	grab4	grab5	grab6
grab2	0.5000000					
grab3	0.7142857	1.0000000				
grab4	0.4285714	0.6666667	0.6666667			
grab5	0.6250000	0.8571429	0.6666667	0.5714286		
grab6	0.5714286	0.6000000	1.0000000	0.7142857	0.5000000	
grab7	0.6666667	0.8750000	0.5000000	0.6250000	0.6250000	0.7500000



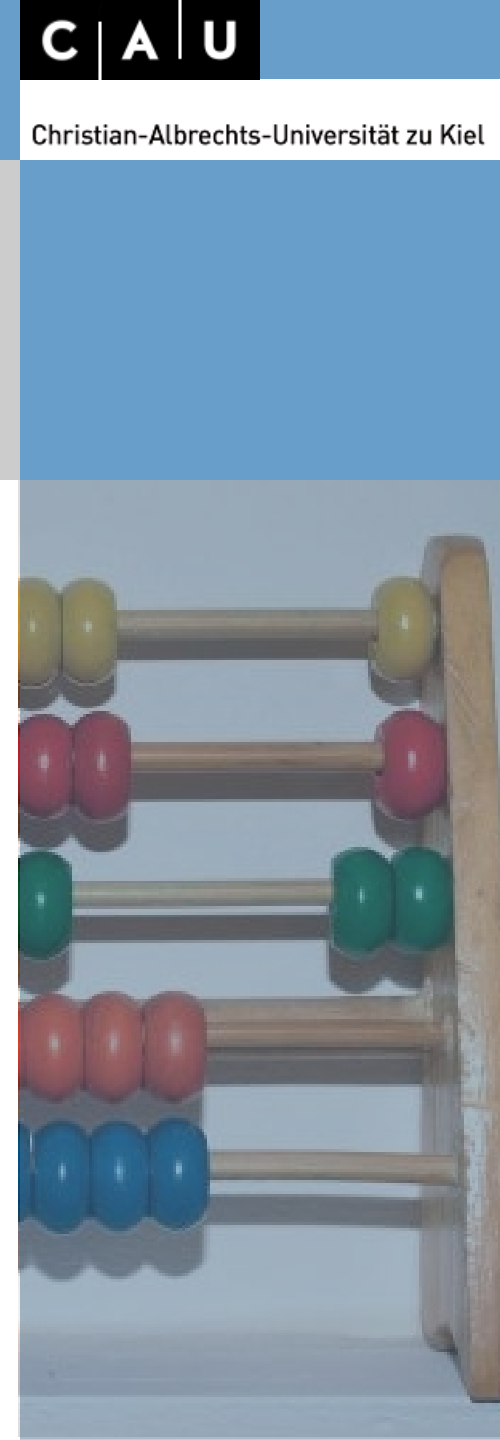
Distanzberechnungen: Aufgabe

Inventare von Siedlungen

Gegeben sind die Inventare verschiedener (hypothetischer) Siedlungen.

Berechnen Sie die passende Distanzmatrix.

Datei: inv_settlement.csv



Distanzberechnungen: Lösung

Inventare von Siedlungen

Gegeben sind die Inventare verschiedener (hypothetischer) Siedlungen.

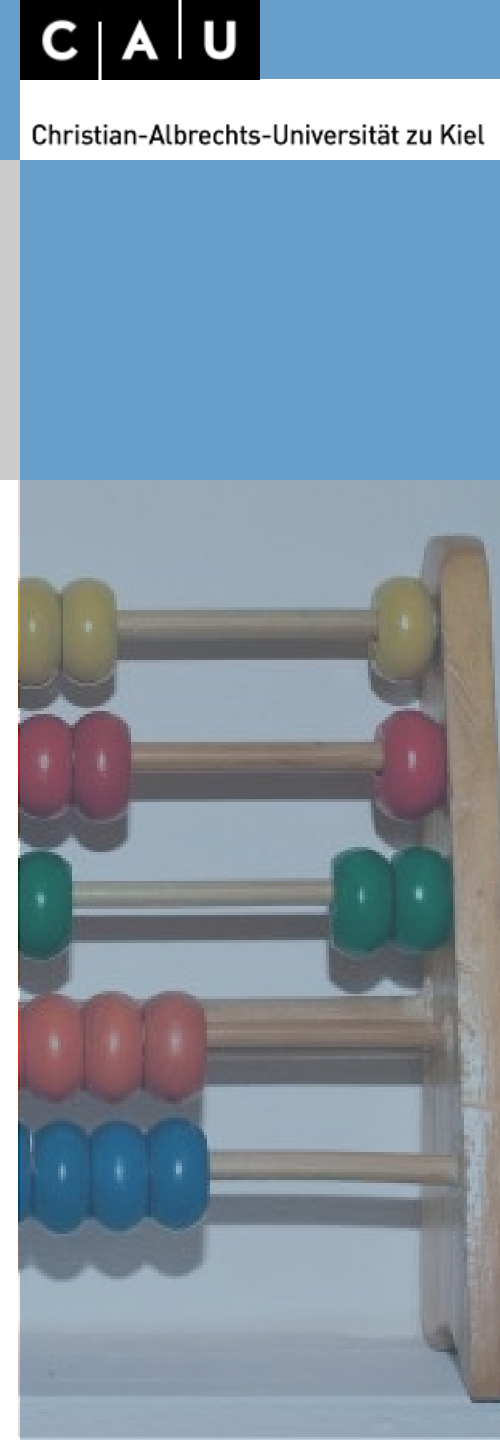
Berechnen Sie die passende Distanzmatrix.

Datei: inv_settlement.csv

```
> inv<-read.csv2("inv_settlement.csv",row.names=1)  
> inv
```

Es handelt sich um Zählwerte, daher Verhältnis-(Absolut-)skala, euklidische Distanz ist möglich

```
> inv.dist<-dist(inv)
```



Hierarchisches Clustern [1]

Wenn wir dann die Entfernungen haben...

Beispiel Backhaus et al.: Magarine

Euklidische Distanz
Matrix, Berechnet aus
Versch. Faktoren

Am Ähnlichsten sind:

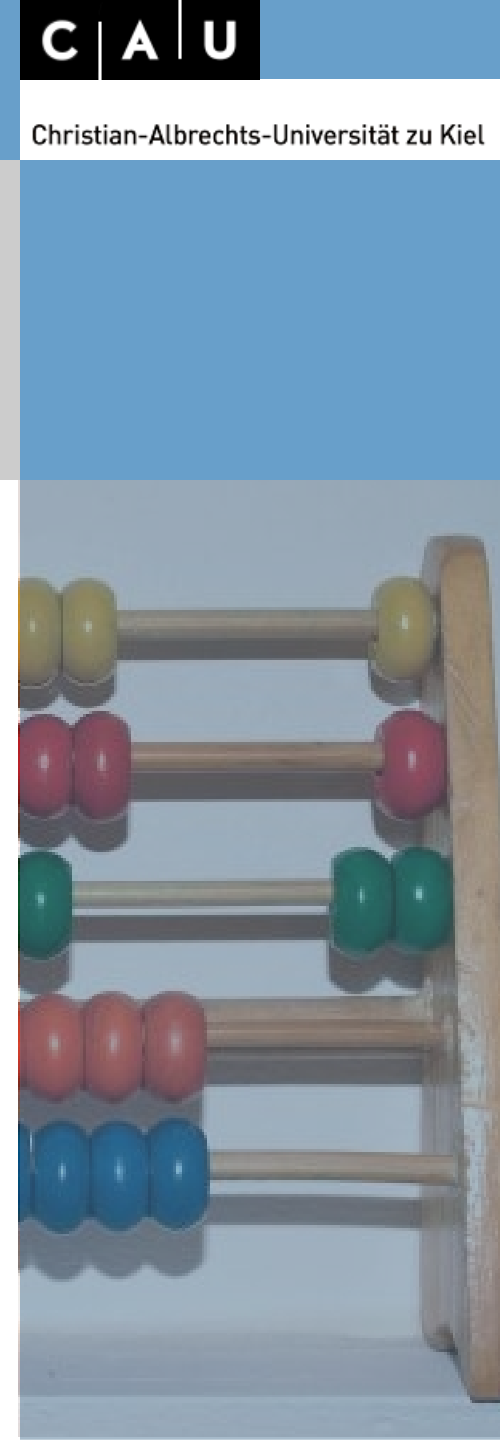
Flora und Rama.

	Rama	Homa	Flora	SB
Homa	6			
Flora	4	6		
SB	56	26	44	
Weihnachts butter	75	41	59	11

Diese bilden unser erstes Cluster bei einer Distanz von 4

Für die weiteren Schritte gibt es verschiedene Verfahren, um den Wert für den neuen Cluster zu bestimmen...

Clusterung bei: {4}



Hierarchisches Clustern [2]

Position von Clustern, Methoden

Single Linkage-Verfahren

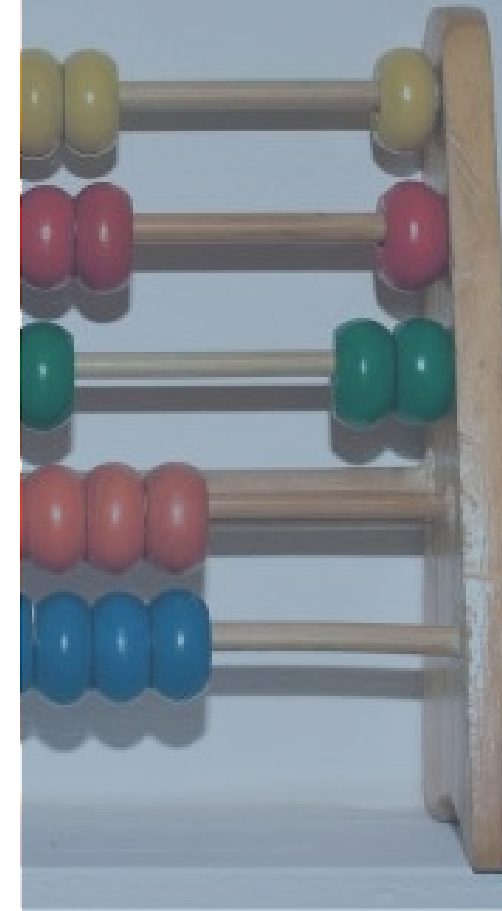
Nächstgelegener Nachbar: Die Abstand von der Gruppe {Rama,Flora} bestimmt sich aus dem kleinsten Abstand von dieser Gruppe zu allen anderen Werten.

Bsp:

	Rama	Homa	Flora	SB
Homa	6			
Flora	4	6		
SB	56	26	44	
Weihnachtsbutter	75	41	59	11

	Rama, Flora	Homa	SB
Homa	6		
SB	44	26	
Weihnachtsbutter	59	41	11

Clusterung bei {4}



Hierarchisches Clustern [3]

Position von Clustern, Methoden

Single Linkage-Verfahren

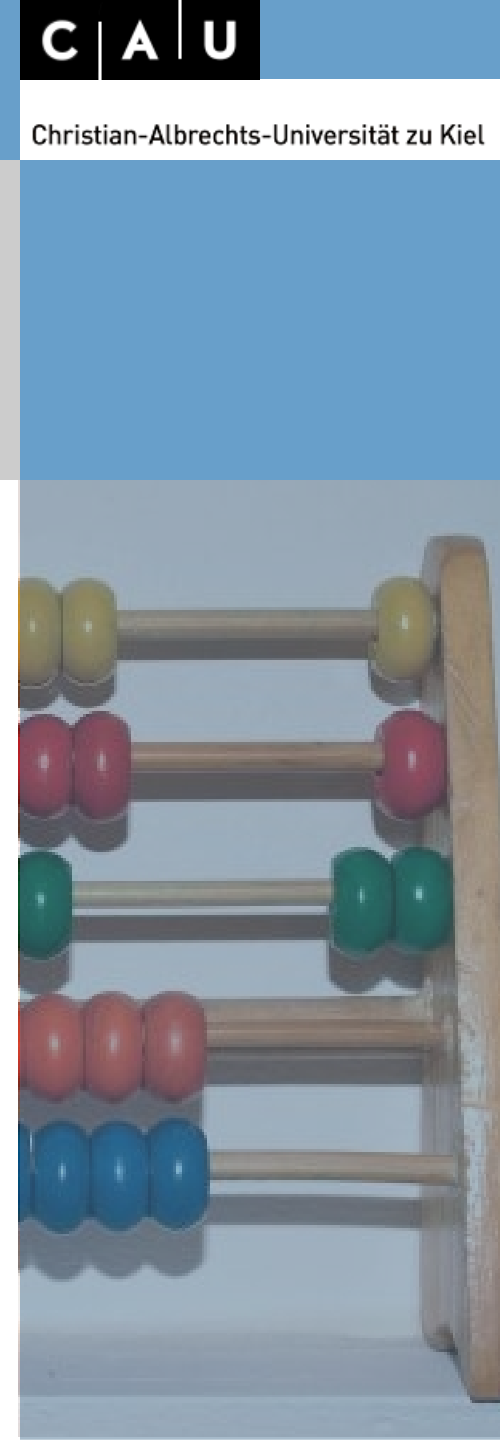
Nächstgelegener Nachbar: Die Abstand von der Gruppe {Rama,Flora} bestimmt sich aus dem kleinsten Abstand von dieser Gruppe zu allen anderen Werten.

Bsp:

	Rama, Flora, Homa	SB
SB	26	
Weihnachts butter	41	11

	Rama, Flora	Homa	SB
Homa	6		
SB	44	26	
Weihnachts butter	59	41	11

Clusterung bei
{4,6}



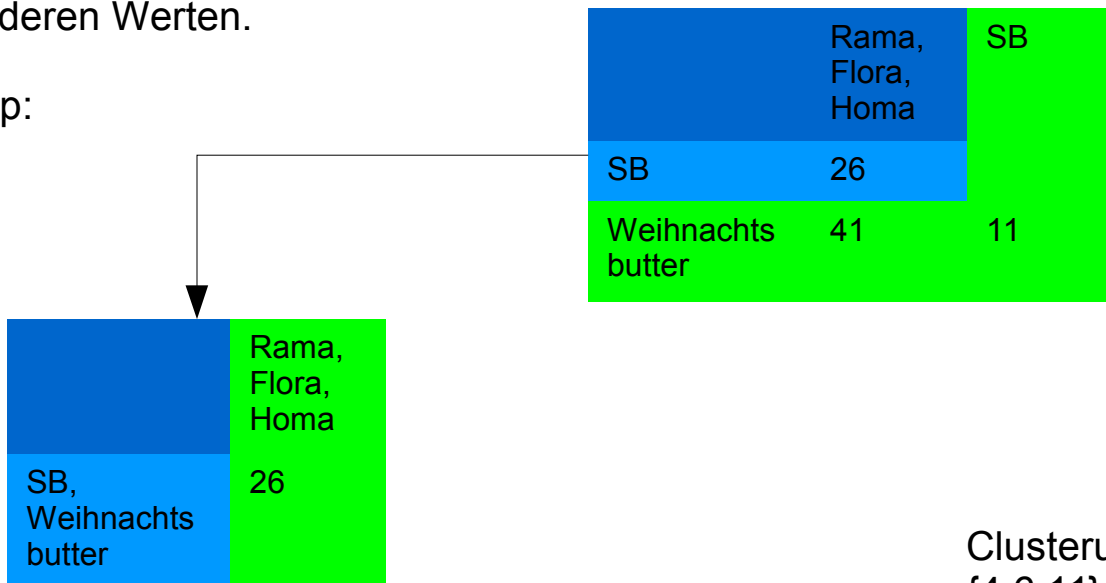
Hierarchisches Clustern [4]

Position von Clustern, Methoden

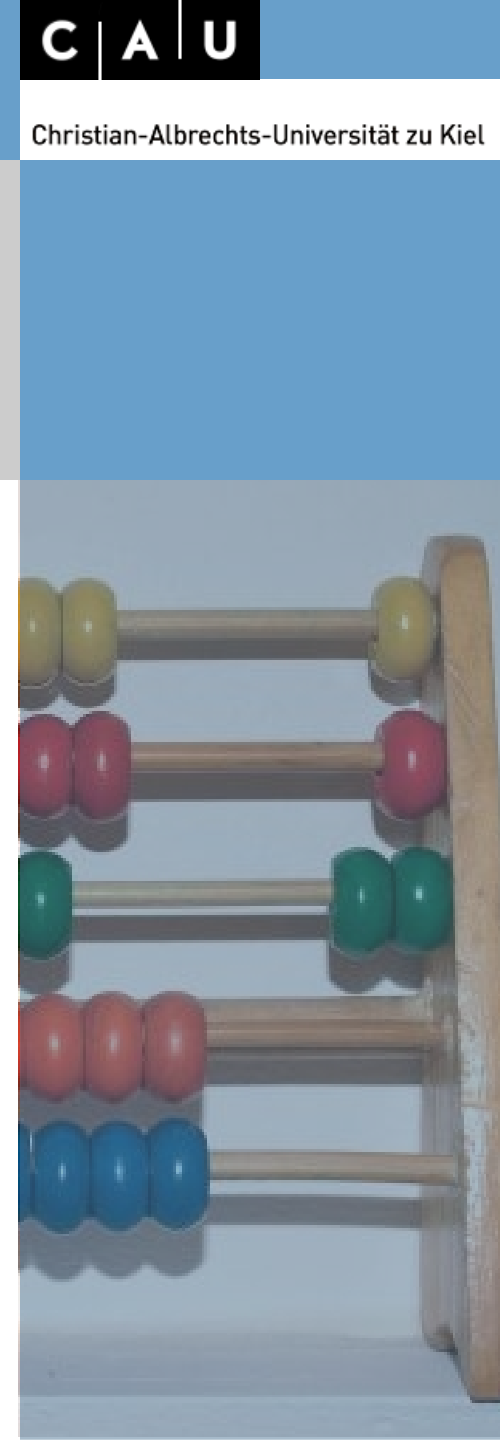
Single Linkage-Verfahren

Nächstgelegener Nachbar: Die Abstand von der Gruppe {Rama,Flora} bestimmt sich aus dem kleinsten Abstand von dieser Gruppe zu allen anderen Werten.

Bsp:



Clustering bei
 $\{4,6,11\} \rightarrow$
 $\{4,6,11,26\}$



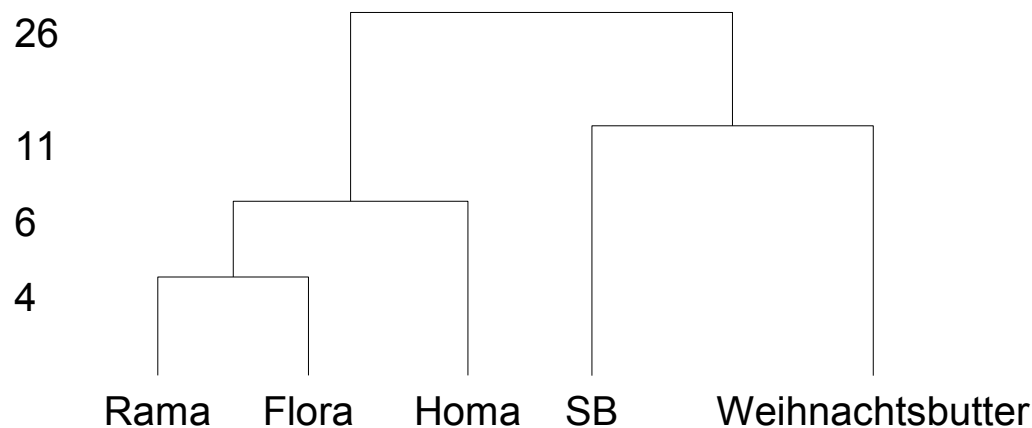
Hierarchisches Clustern [4]

Position von Clustern, Methoden

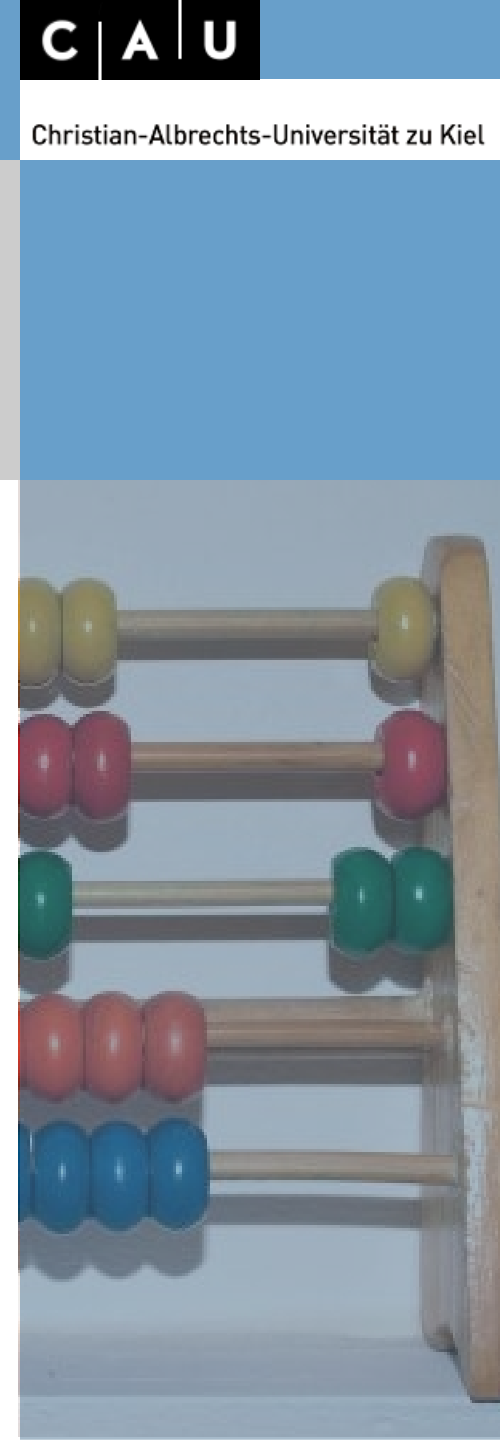
Dendrogramm

Darstellung des Verlaufes der Verbindungen der Cluster untereinander

Bsp:



Clustering bei
 $\{4, 6, 11, 26\}$



Hierarchisches Clustern: Methoden

Andere Methoden

Complete Linkage-Verfahren:

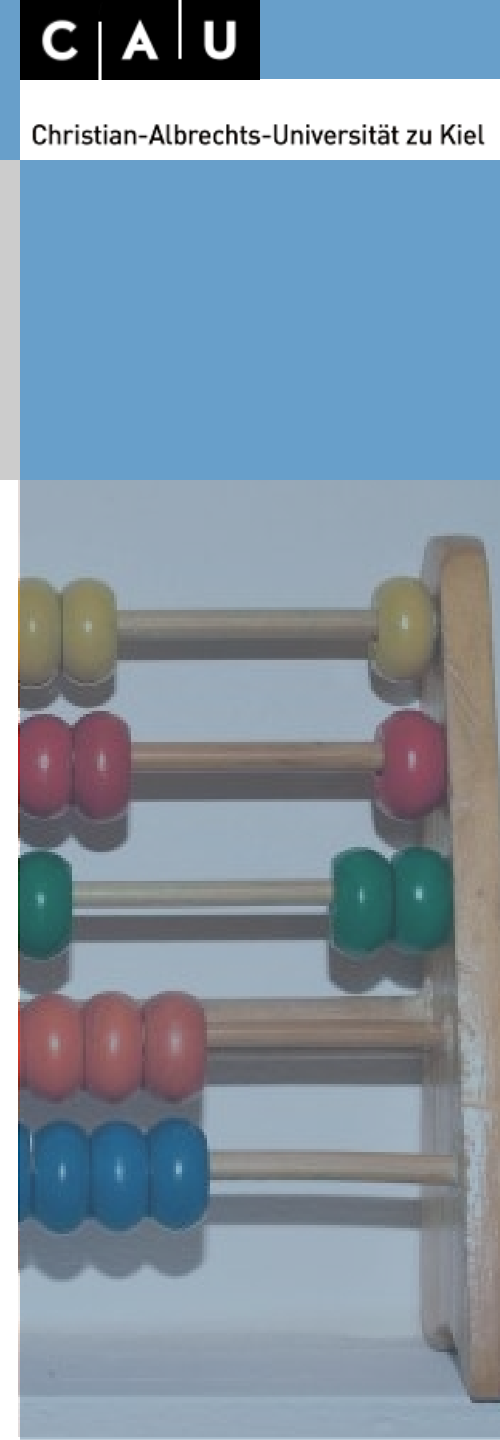
Der am weitesten entfernte Nachbar wird gewählt

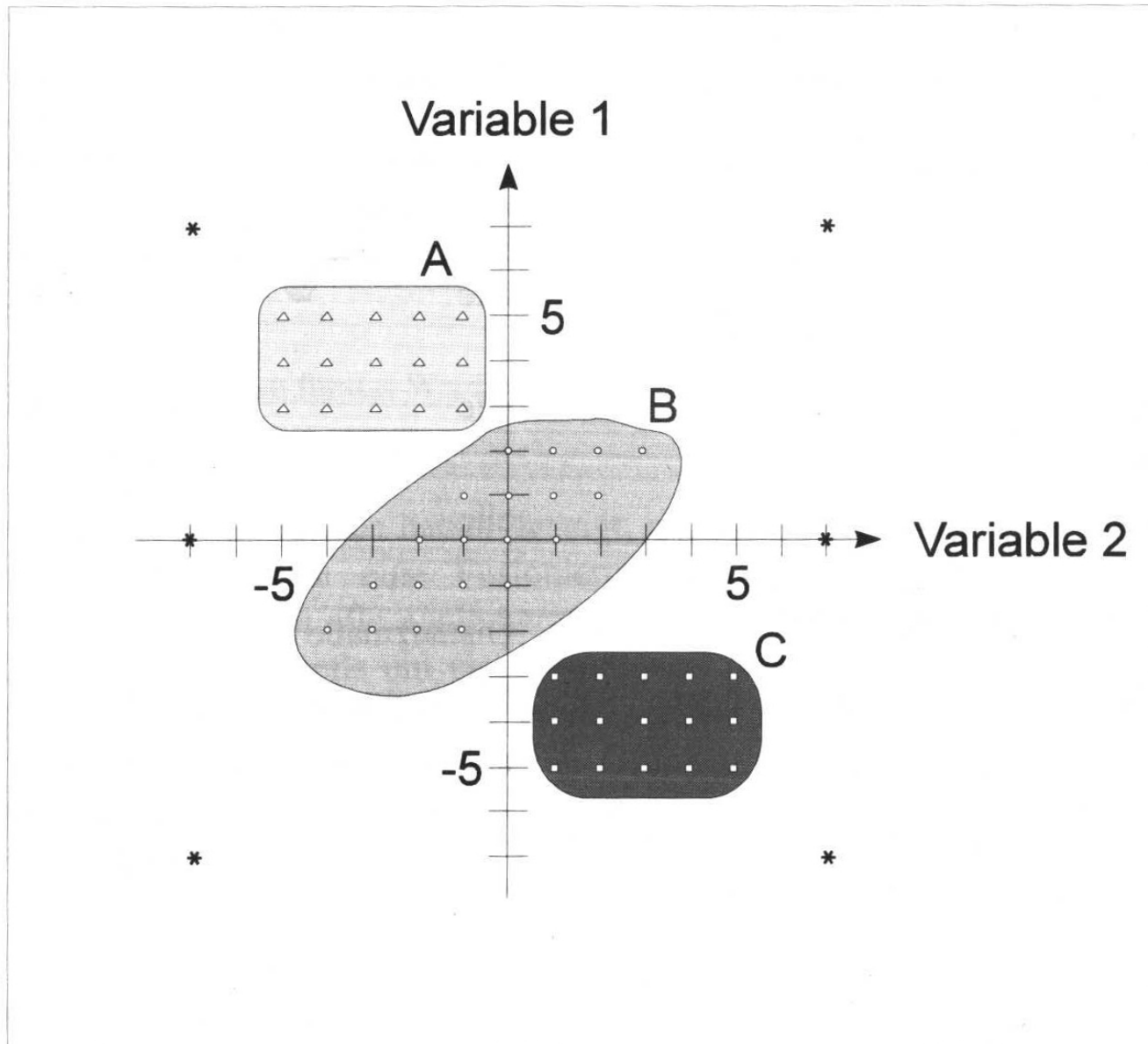
Average Linkage-Verfahren

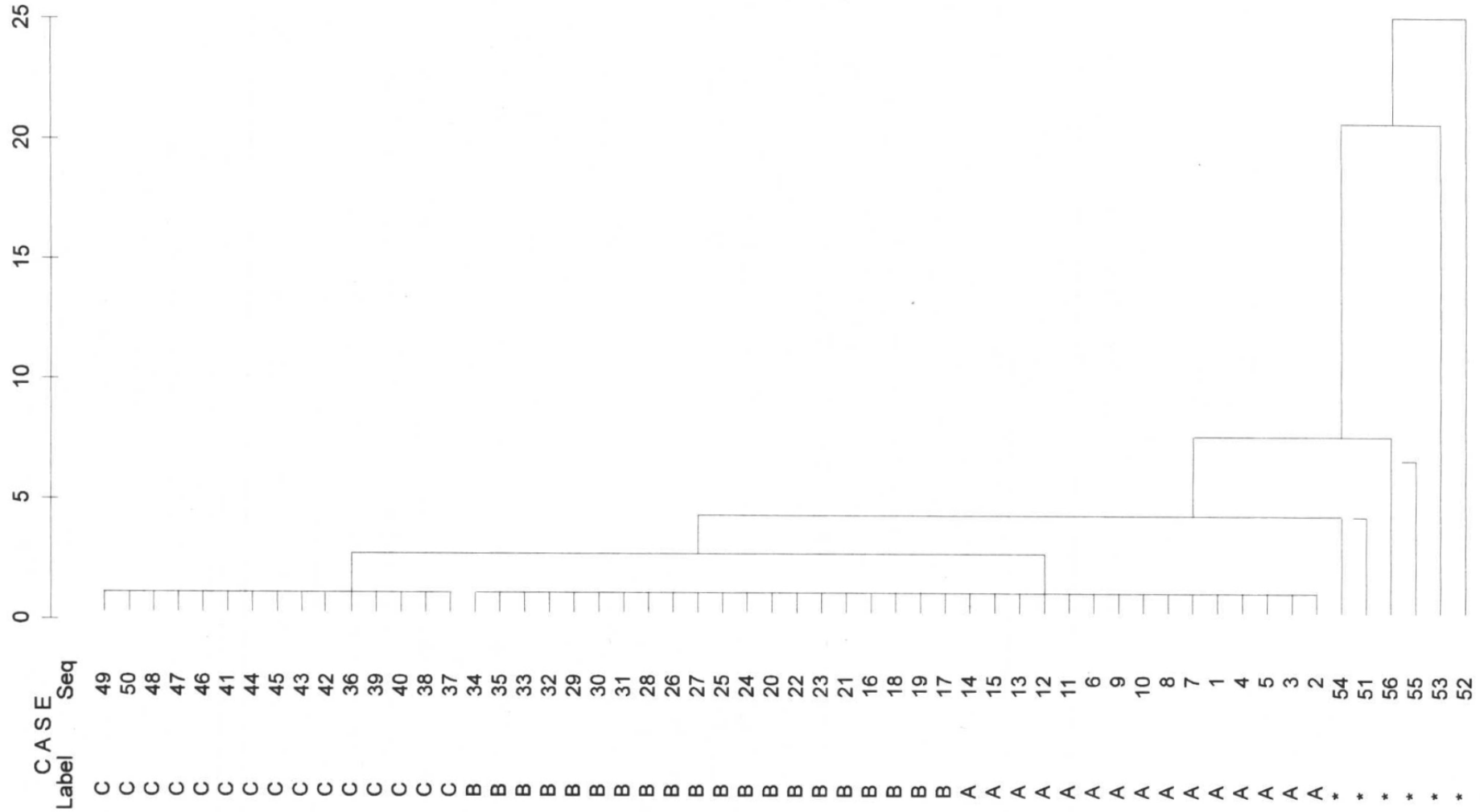
Der Mittelwert der paarweisen Distanzen aller Daten wird gewählt

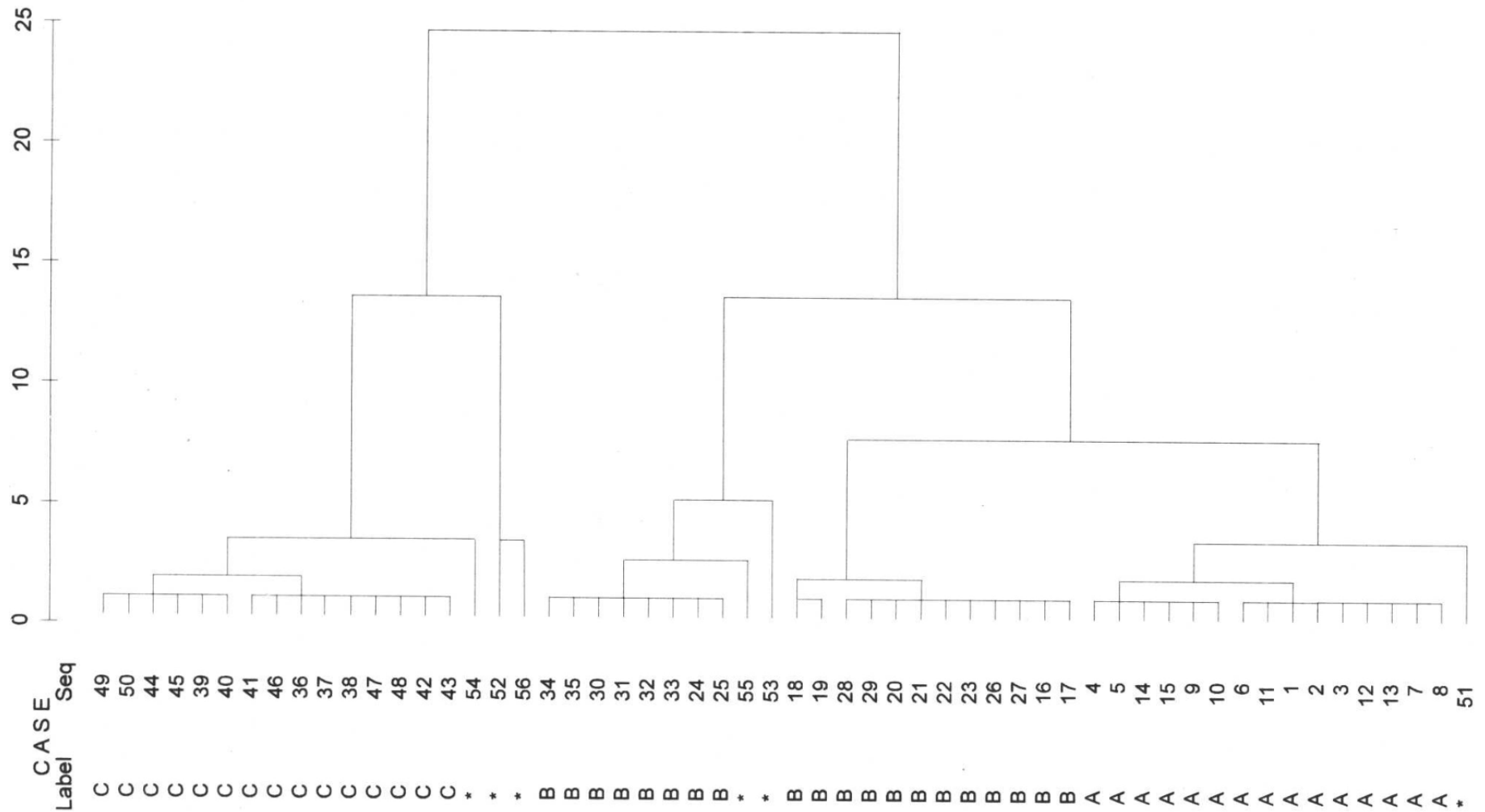
Ward-Verfahren

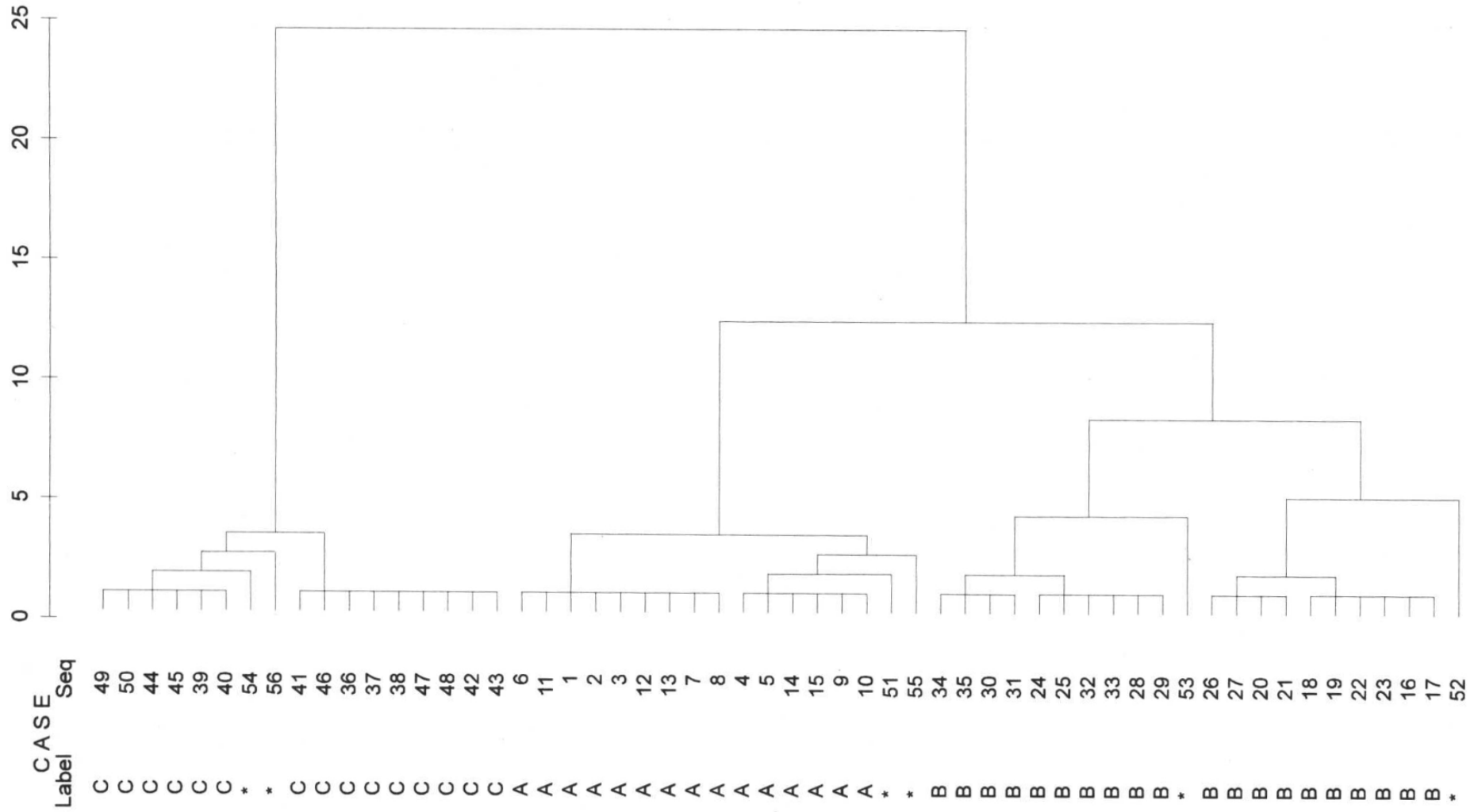
Diejenigen Gruppen werden vereinigt, bei denen durch die Zusammenlegung sich die Varianz innerhalb der Gruppe am wenigsten erhöht. Gutes (Bestes?) Verfahren für die Bestimmung von Clustern, wenn Distanzmaße (metrische Variablen) vorliegen.











Hierarchisches Clustern: Ward-Methode

Verfahren, wenn metrische Daten vorliegen

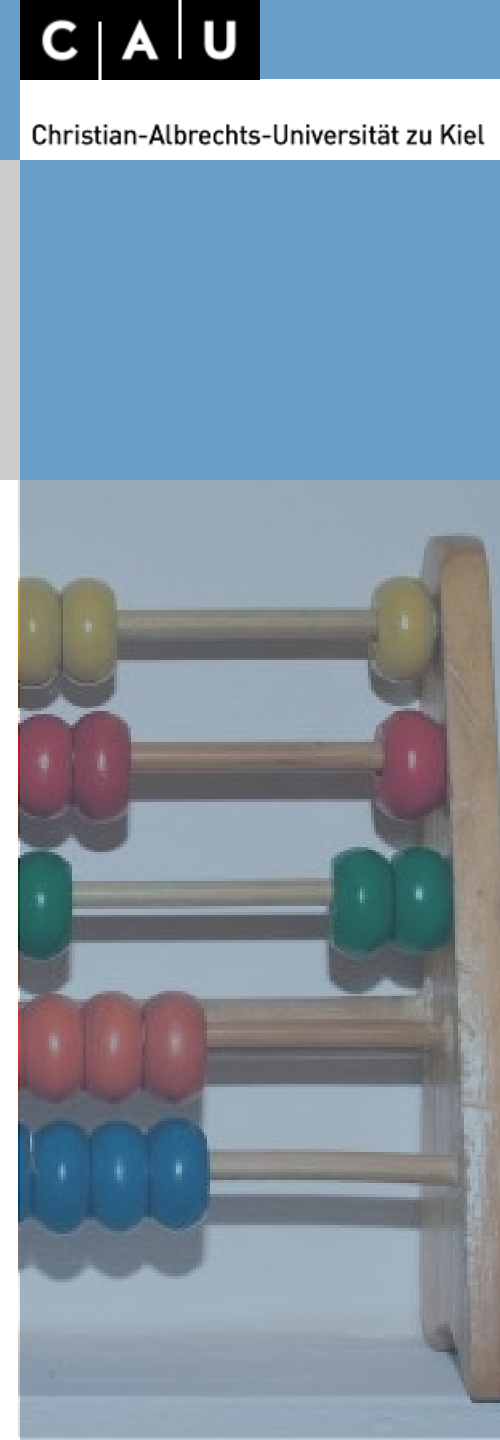
Derjenige Wert wird zu einem Cluster hinzugezogen, der die wenigste Erhöhung der Varianz innerhalb des Clusters bewirkt.

Vorteil: findet meist „natürliche“ Gruppierungen am besten.

Nachteil: ist (von Haus aus) nur für metrisch skalierte Variablen anwendbar [aber: Jaccard-Distanz kann verarbeitet werden]
Findet schlecht Gruppen mit kleiner Elementzahl oder langgestreckte Gruppen

In R:

```
> leder.dist<-dist(leder[,c("Breite","Laenge","Dicke")],method="euclid")  
> leder.clust<-hclust(leder.dist,method="ward")  
> plot(leder.clust)
```



Hierarchisches Clustern: Average-Linkage-Methode

Ein Verfahren, wenn nominale Daten vorliegen

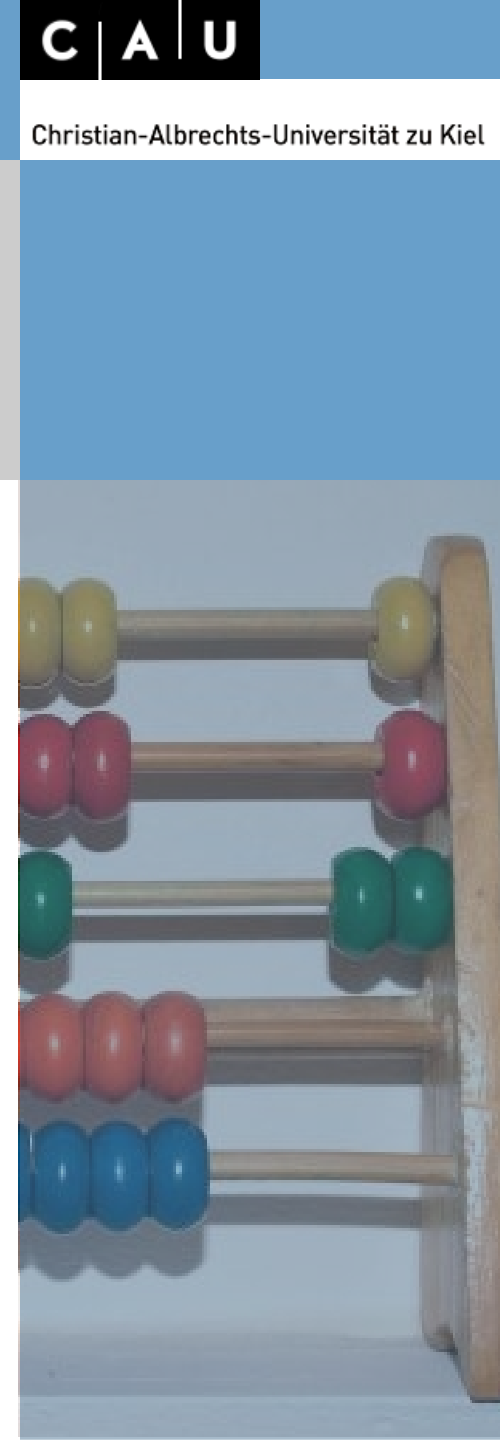
Das neue Distanzmaß wird aus dem Durchschnitt aller paarweisen Vergleiche der Distanzen der Mitglieder zweier Cluster errechnet

Vorteil: läßt sich auch bei nominal skalierten Variablen anwenden, berücksichtigt alle Elemente eines Clusters bei der Neubestimmung der Distanzen

Nachteil: Nicht so gut geeignet wie Ward, um „natürliche“ Gruppen zu finden

In R:

```
> graeber.dist<-vegdist(graeber,method="jaccard")  
> graeber.clust<-hclust(graeber.dist,method="average")  
> plot(graeber.clust)
```



Hierarchisches Clustern: Clusteranzahl

Wieviel Gruppen reichen?

1. Inhaltliche Überlegungen

Wieviele Gruppen erwarte ich? Sind sinnvoll? Lassen sich aus dem Dendrogramm ablesen?

2. Elbow-Kriterium

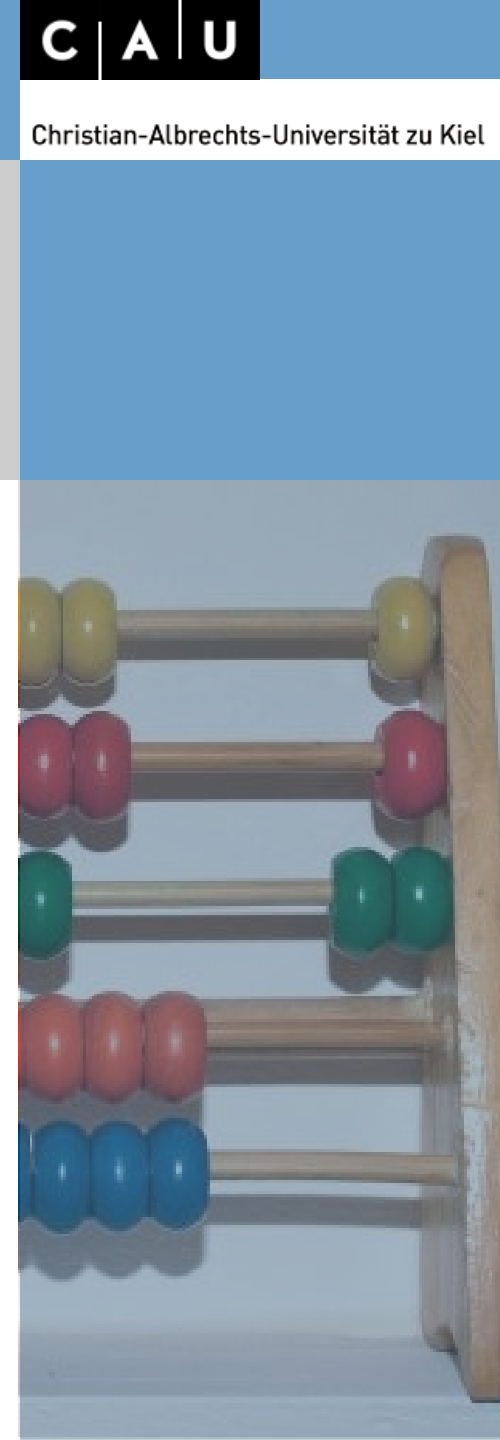
Für Ward-Clustering: Wenn sich die Varianz innerhalb der Clusters nicht mehr stark erhöht, ist eine gute Clusterung gefunden.

In R:

Anzeige für die letzten 10 Cluster

```
> plot(rev(leder.clust$height) [1:10] , type="l")
```

„Ellenbogen“ bei 5: 5 Cluster-Lösung scheint sinnvoll



Hierarchisches Clustern: Darstellung

In 2/3 Dimensionen

Dendrogramm

```
> plot(leder.clust)
```

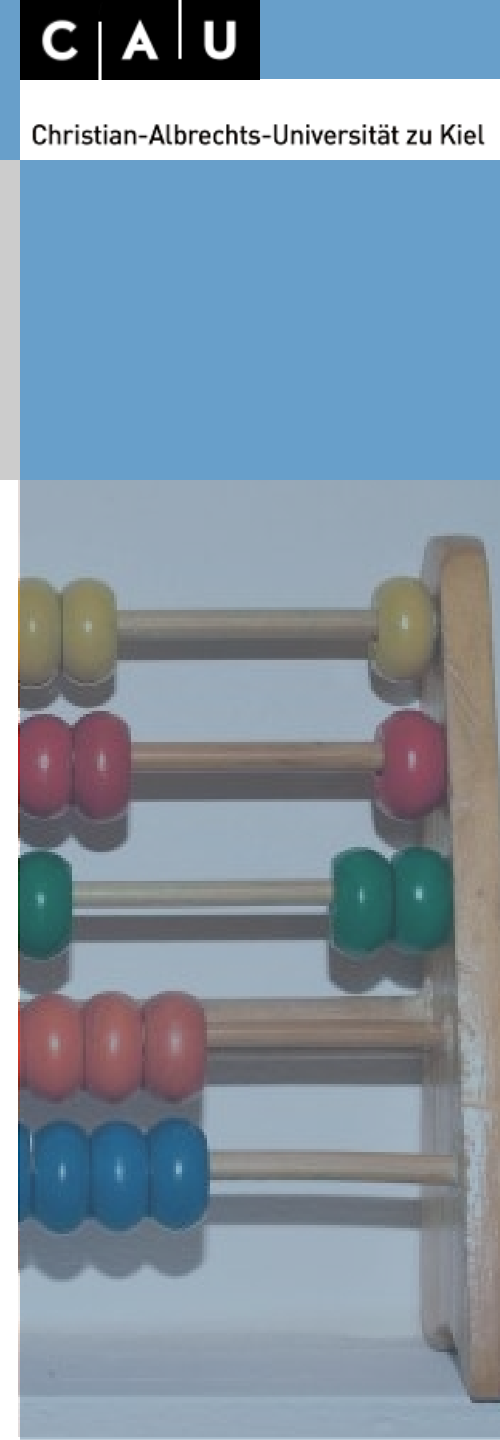
In einem 2D-Diagramm

Für die 5-Cluster Lösung, unterschiedliche Cluster werden unterschiedlich eingefärbt

```
> leder.clusters.5<-cutree(leder.clust,5)  
> plot(leder$Laenge,leder$Breite,col=rainbow(5)[leder.clusters.5])
```

In einem 3D-Diagramm

```
> library(lattice)  
> cloud(leder$Breite~leder$Laenge+leder$Dicke,col=rainbow(5)  
[leder.clusters.5])
```



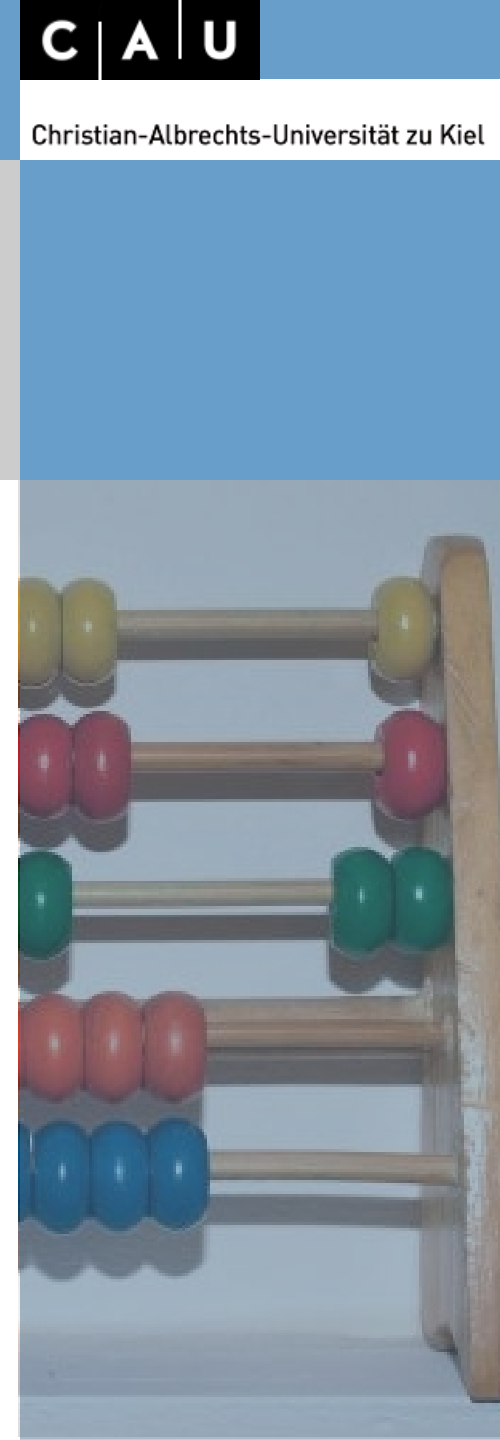
Hierarchisches Clustern: Aufgabe

Keramik mit verschiedenen Dekorationselementen

Gegeben sind Keramikfunde mit verschiedenen Eigenschaften. Bestimmen Sie, welches Distanzmaß angebracht ist, berechnen Sie die Distanzmatrix und führen Sie eine Clusteranalyse mit einer passenden Methode aus.

Bestimmen Sie eine gute Clusterlösung und zeigen Sie das Dendrogramm an.

File: ceramics.csv



Hierarchisches Clustern: Aufgabe

Keramik mit verschiedenen Dekorationselementen

Gegeben sind Keramikfunde mit verschiedenen Eigenschaften. Bestimmen Sie, welches Distanzmaß angebracht ist, berechnen Sie die Distanzmatrix und führen Sie eine Clusteranalyse mit einer passenden Methode aus. Bestimmen Sie eine gute Clusterlösung und zeigen Sie das Dendrogramm an.

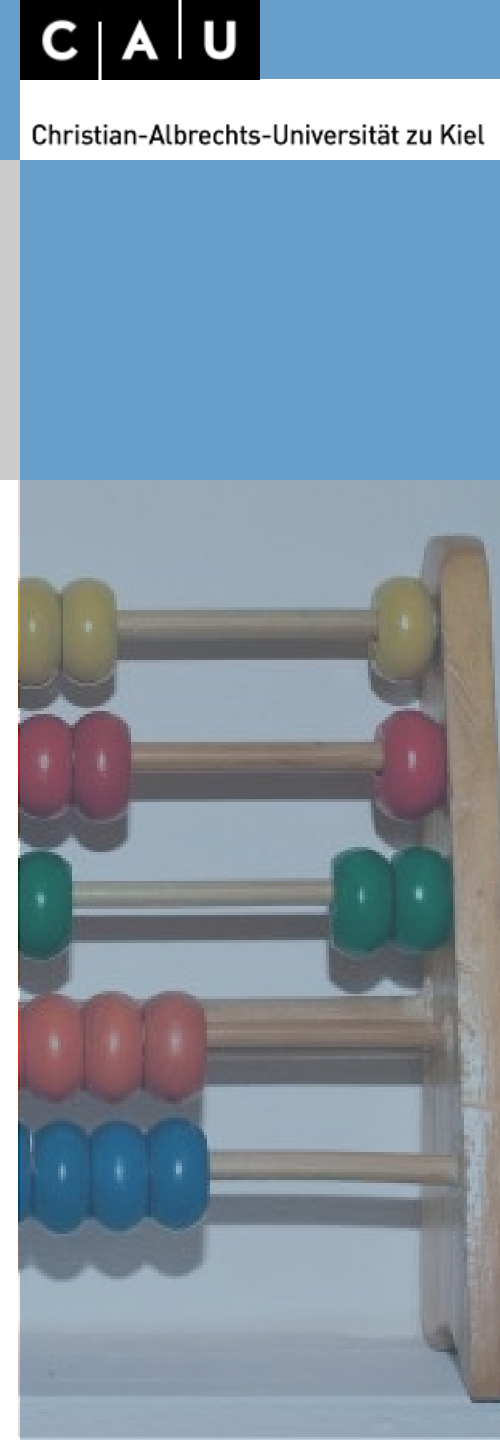
File: ceramics.csv

```
> ceramics<-read.csv2("ceramics.csv",row.names=1)
> ceramics
```

Es handelt sich um eine Präsenz-Absenz-Matrix, daher Distanz Jaccard, Clustermethode z.B. Average. Ein Ellenbogenkriterium kann nicht angewandt werden

```
> ceramics.dist<-vegdist(ceramics,method="jaccard")
> ceramics.clust<-hclust(ceramics.dist,method="average")
> plot(ceramics.clust)
```

Es bilden sich deutlich zwei Cluster von Gefäßen.



Nichthierarchisches Clustern [1]

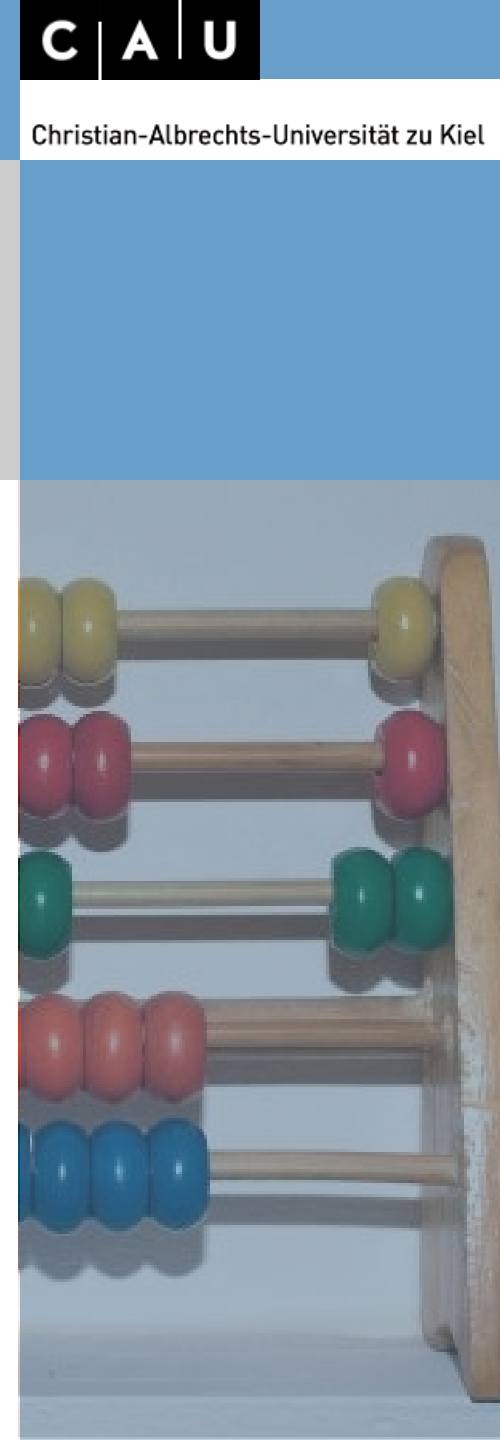
Wenn man eine Clusterzahl voraussetzen kann...

In jedem Schritt werden die Cluster neu zusammengestellt und neue Anstände berechnet. Wenn die Lösung möglichst optimal ist, bricht das Verfahren ab.

Bsp. kmeans:

Mögliches Vorgehen: identifizieren der optimalen Clusterzahl mit hierarchischer Methode (Ward), dann eigentliches Clustern mit kmeans

Daten: andean_sites.csv



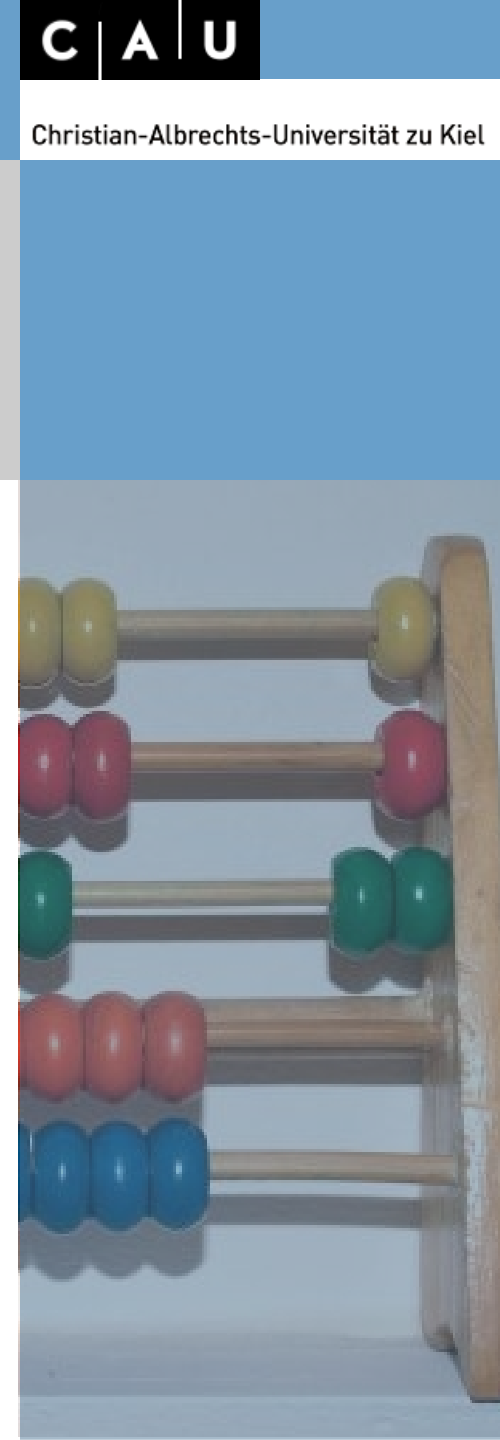
Nichthierarchisches Clustern [2]

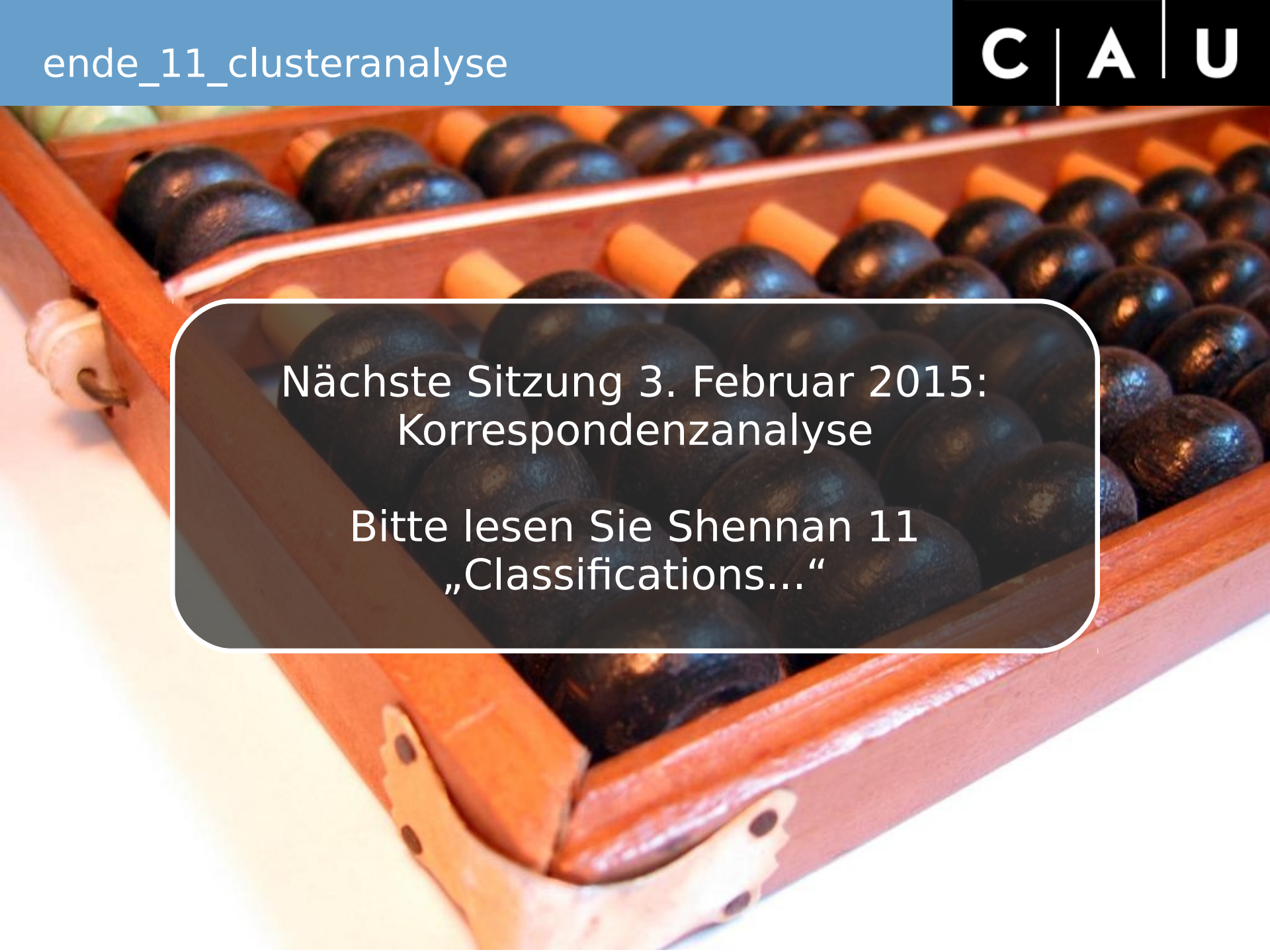
Wenn man eine Clusterzahl voraussetzen kann...

```
> andean<-read.csv2("andean_sites.csv",row.names=1)
> andean.hclust<-hclust(dist(andean),method="ward")
> plot(rev(andean.hclust$height),type="l")
```

Ellenbogen bei 3, daher 3 Cluster

```
> andean.kmeans<-kmeans(andean,3)
> andean.kmeans
...
> plot(andean,col=andean.kmeans$cluster)
```





Nächste Sitzung 3. Februar 2015:
Korrespondenzanalyse

Bitte lesen Sie Shennan 11
„Classifications...”