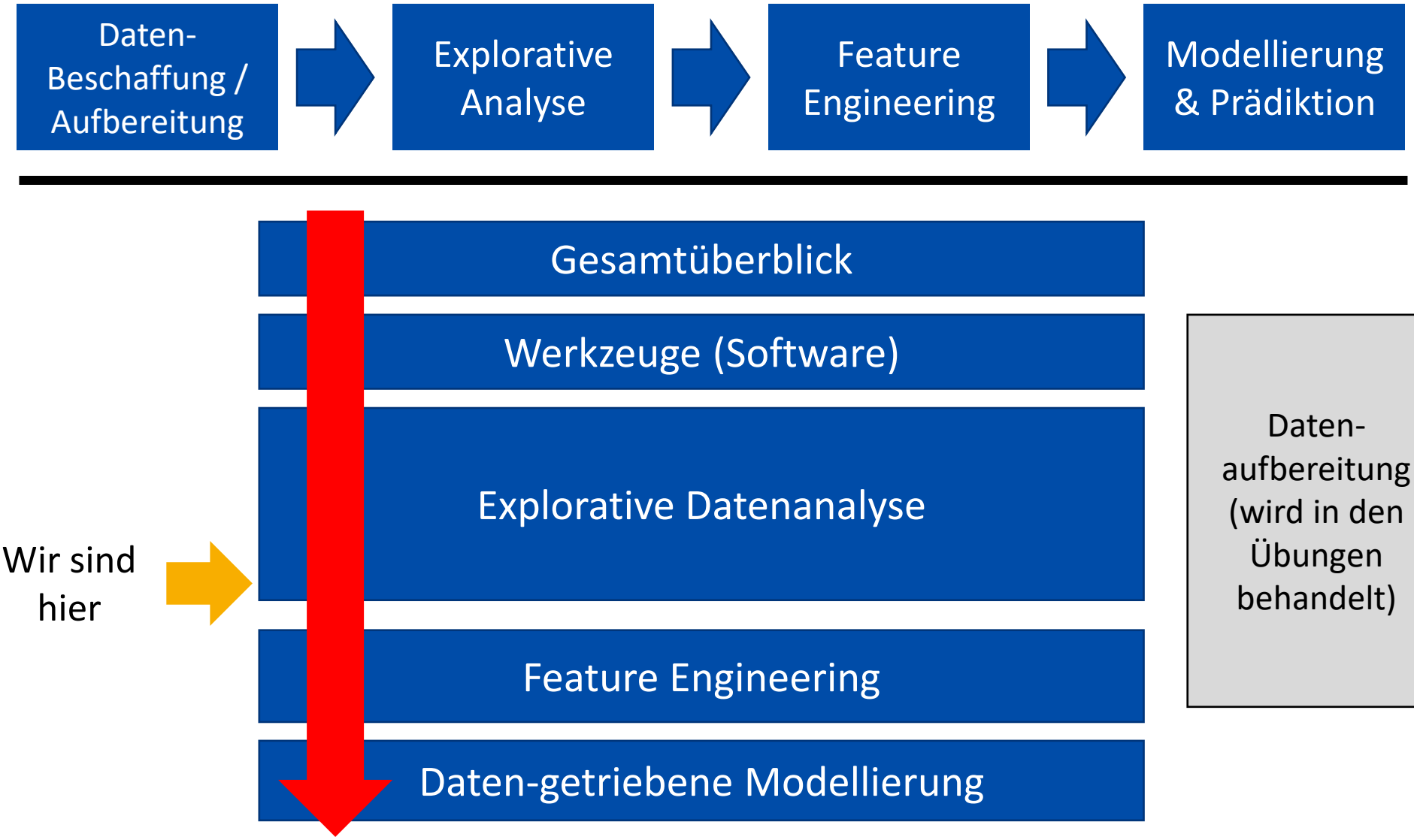
Einführung in Data Science

Unser Plan für diese Vorlesungseinheit:

Clustering

1. K-Means Clustering
2. Hierarchische Clusteranalyse (HCA) (*hierarchical clustering*)

Data Science



Curriculum

1. Grundbegriffe / Überblick
2. Zentrale Softwarebibliotheken

3. Univariate explorative Analyse (EDA) Visualisierung (Teil 1)
4. Visualisierung (Teil 2), Multivariate explorative Analyse (Teil 1)
5. Multivariate explorative Analyse (Teil 2)
6. Dimensionsreduktion (Teil 1): PCA
7. Dimensionsreduktion (Teil 2): MDS, Isomap
8. Clustering: K-Means, HCA
9. Clustervalidierung

10. Probeklausur

11. Feature Engineering, Datengetriebene Modellierung (Teil 1)
12. Datengetriebene Modellierung (Teil 2)

Überblick / Begriffe

Explorative Analyse (EDA)

Feature Engineering & Modellierung

Multivariate Explorative Analyse

Methoden der multivariaten explorativen Analyse:

1. Multivariate deskriptive Statistik
   (hier vor allem: Visualisierungsarten)

2. Korrelationskoeffizienten
   (bivariate Analyse: Suche nach Zusammenhängen)

3. Dimensionsreduktionsverfahren

4. Clusteranalyse

Clusteranalyse

Clusteranalyse

- Methoden, die die Identifizierung von Gruppen (Cluster) *ähnlicher* Datenpunkte in einem Datensatz ermöglichen

Clustering: Partition (Einteilung) der Datenpunkte in Gruppen

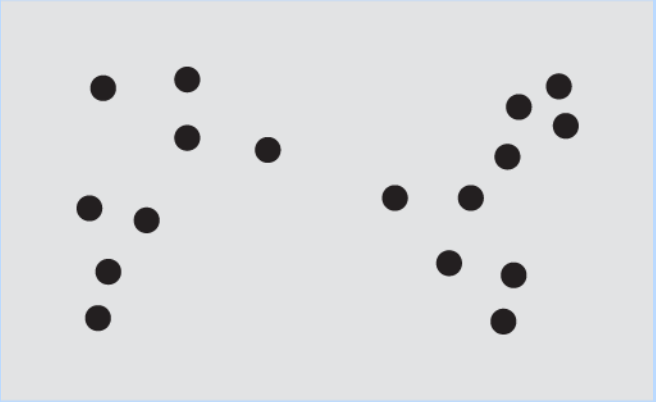
Viele Cluster-Analyseverfahren existieren und unterscheiden sich

- in der Definition von Ähnlichkeit
- in der Definition von Gruppen (Clustern)
- im algorithmischen Vorgehen

Clustering

Interaktivität F

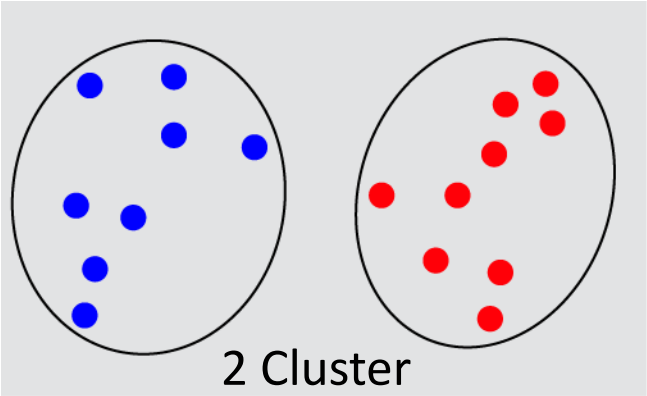
Gegeben ist der rechts dargestellte 2-dimensionale Datensatz.

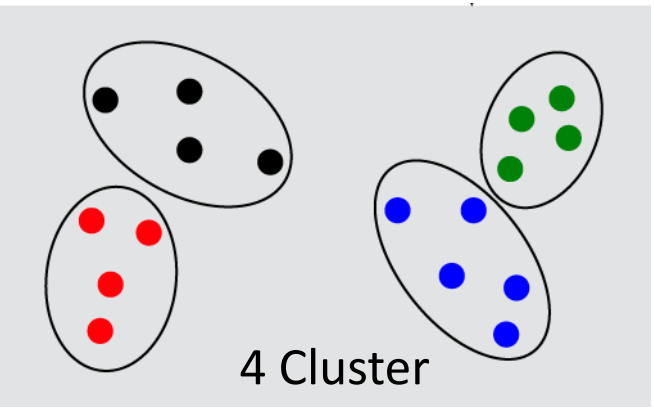


1. Entscheiden Sie mit Ihrem Banknachbarn, wie viele und welche Cluster der Datensatz enthält.
2. Wenn Sie mehrere Clusterings finden: Welches Clustering ist richtig? Welches Clustering ist falsch?

Zu 1): Beispiele

Zu 2): Nicht direkt entscheidbar. „Es kommt drauf an.“





Clustering

Ergebnisse:

1. Cluster-Analyseverfahren finden immer Cluster in den Daten.
2. Verschiedene Verfahren liefern verschiedene Cluster für dieselben Daten.
3. Cluster-Verfahren interpretieren die Daten hinsichtlich der den Verfahren (teils implizit) zugrundeliegenden Annahmen.
4. Wie Partitionen bewertet werden können (also z.B. als sinnvoll/nicht sinnvoll eingestuft werden können), ist aktuelles und im Allgemeinen ungelöstes Forschungsproblem.

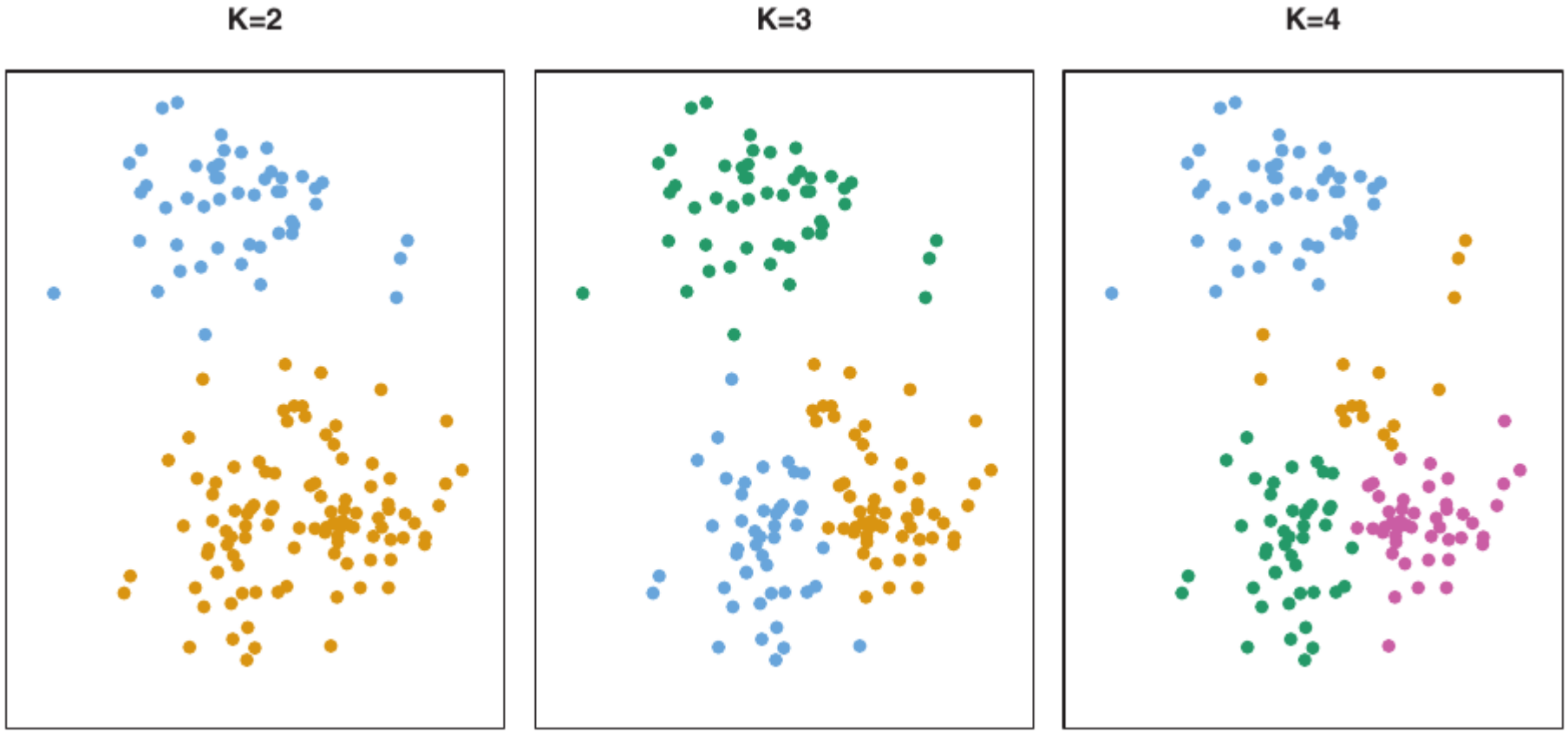
Wir betrachten im Folgenden zwei Cluster-Verfahren:

- K-Means Clustering
- Hierarchisches Clustering

K-Means Clustering

- findet K nicht überlappende Cluster
- Anzahl K der Cluster muss vom Nutzer angegeben werden

Beispiel:



K-Means Clustering

Seien $C_1, \ldots, C_K$ die Mengen der Indizes der Datenpunkte jedes Clusters.

Die Mengen sollen zwei Eigenschaften genügen:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = 1, \ldots, N$
   Auf Deutsch: Jeder Datenpunkt gehört zu mindestens einem Cluster.

2. $C_k \cap C_{k'} = \emptyset \quad \forall k \neq k'$
   Auf Deutsch: Die Cluster überlappen nicht. Jeder Datenpunkt gehört nicht mehr als einem Cluster an.

Beispiel
Wenn der *i*-te Datenpunkt zum *k*-ten Cluster gehört, dann gilt:

$$i \in C_k$$

K-Means Clustering

Hauptidee hinter K-Means Clustering:

- Clustering ist gut, wenn die *Intra-Cluster-Variation* minimiert wird.

Bezeichne $W(C_k)$ die Intra-Cluster-Variation des Clusters C_k.

Wir formalisieren die obige Idee:

$$\underset{C_1,\ldots,C_K}{\text{minimiere}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Die Summe der Intra-Cluster-Variation aller Cluster sei also minimal.

Intra-Cluster-Variation

- wird meist definiert als Summe der *quadratischen euklidischen Distanzen* zwischen allen Datenpunkten eines Clusters

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Anzahl Datenpunkte im Cluster k

selektiert Datenpunkte aus Cluster k

j-te Komponente des Datenpunktes (Featurevektors) i

K-Means Clustering

Einsetzen der zwei mathematischen Ausdrücke liefert das folgende

Optimierungsproblem – *k*-means Clustering
(*Clustering durch Varianzminimierung*)

$$\underset{C_1,\ldots,C_K}{\text{minimiere}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

Optimierungsfunktion (Kostenfunktion)

Optimierungsproblem ist schwer zu lösen, daher arbeiten wir (und alle anderen) mit einem **Algorithmus**, der eine (nicht notwendigerweise die beste) Lösung liefert.

K-Means Clustering

Algorithmus

1. Weise allen Datenpunkten zufällige Ziffern 1, …, *K* zu.
   Sie dienen als zufällige Clusterzugehörigkeiten.

2. Iteriere, bis die Cluster sich nicht mehr ändern:

   a) Für jeden der K Cluster, berechne den geometrischen Schwerpunkt (*centroid*). Der Schwerpunkt ist der Featurevektor, der sich aus der Mittelung aller Featurevektoren der dem Cluster zugehörigen Datenpunkte ergibt.

   b) Weise jedem Datenpunkt dem Cluster zu, dessen Schwerpunkt dem Datenpunkt am nächsten liegt (im euklidischem Sinne).

K-Means Clustering

Optimierungsproblem – *k*-means Clustering
(*Clustering durch Varianzminimierung*)

$$\underset{C_1,\ldots,C_K}{\text{minimiere}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

Frage: Warum verkleinert der Algorithmus obiges Optimierungskriterium?

Es gilt folgende Gleichung:

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 \qquad (1)$$

mit *Centroid*: $\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$

Antwort: In Schritt 2b des Algorithmus wird durch die Neuzuweisung der Datenpunkte die rechte Seite von Gleichung (1) minimiert, und dies minimiert auch die linke Seite von Gleichung (1) (also das Optimierungsproblem).

K-Means Clustering

Beweis von Gleichung (1):

Addition von Null: $-\bar{x}_{kj} + \bar{x}_{kj}$

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} ((x_{ij} - \bar{x}_{kj}) - (x_{i'j} - \bar{x}_{kj}))^2$$

$$= \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} ((x_{ij} - \bar{x}_{kj})^2 - 2(x_{ij} - \bar{x}_{kj})(x_{i'j} - \bar{x}_{kj}) + (x_{i'j} - \bar{x}_{kj})^2)$$

binomische Formel

$$= \frac{|C_k|}{|C_k|} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 + \frac{|C_k|}{|C_k|} \sum_{i' \in C_k} \sum_{j=1}^{p} (x_{i'j} - \bar{x}_{kj})^2$$

Summation über *i* bzw *i´*

$$\underbrace{- \frac{2}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})(x_{i'j} - \bar{x}_{kj})}$$

$$= 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 + 0$$

(*) auf nächster Folie

K-Means Clustering

Zu (*):

Umsortieren der Summen

$$
\begin{aligned}
& -\frac{2}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})(x_{i'j} - \bar{x}_{kj}) \\
& = -\frac{2}{|C_k|} \sum_{j=1}^{p} \sum_{i \in C_k} \sum_{i' \in C_k} (\boxed{x_{ij} x_{i'j}} + \boxed{\bar{x}_{kj}\bar{x}_{kj} - x_{ij}\bar{x}_{kj}} - \boxed{x_{i'j}\bar{x}_{kj}}) \\
& = -\frac{2}{|C_k|} \sum_{j=1}^{p} \sum_{i \in C_k} (|C_k| x_{ij}\bar{x}_{kj} + |C_k|\bar{x}_{kj}\bar{x}_{kj} - |C_k| x_{ij}\bar{x}_{kj} - |C_k|\bar{x}_{kj}\bar{x}_{kj}) \\
& = 0
\end{aligned}
$$

Anwenden von (2)

Anwenden der Summe über *i´*

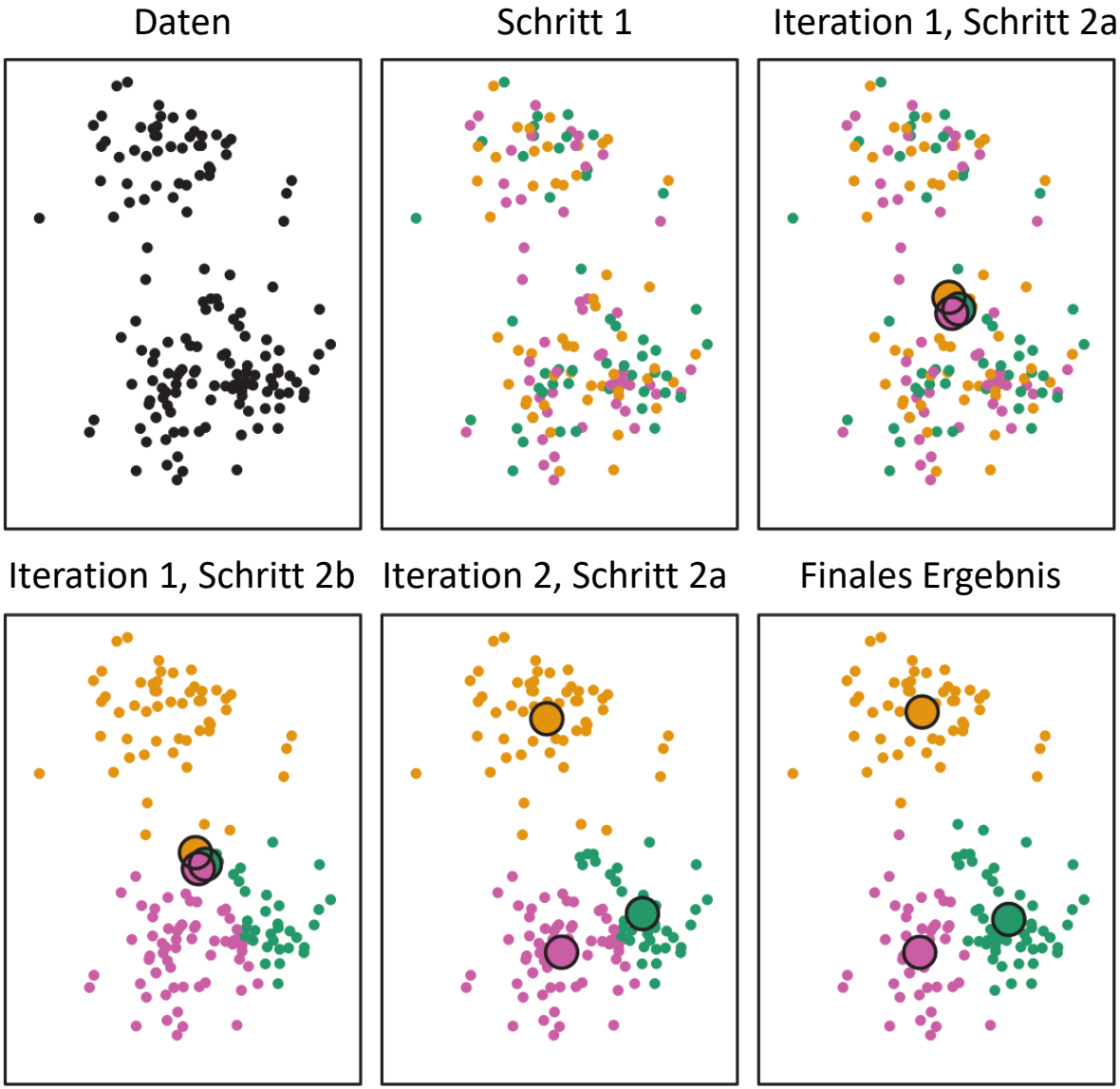
Anwenden von (2)

Definition des Centroids:

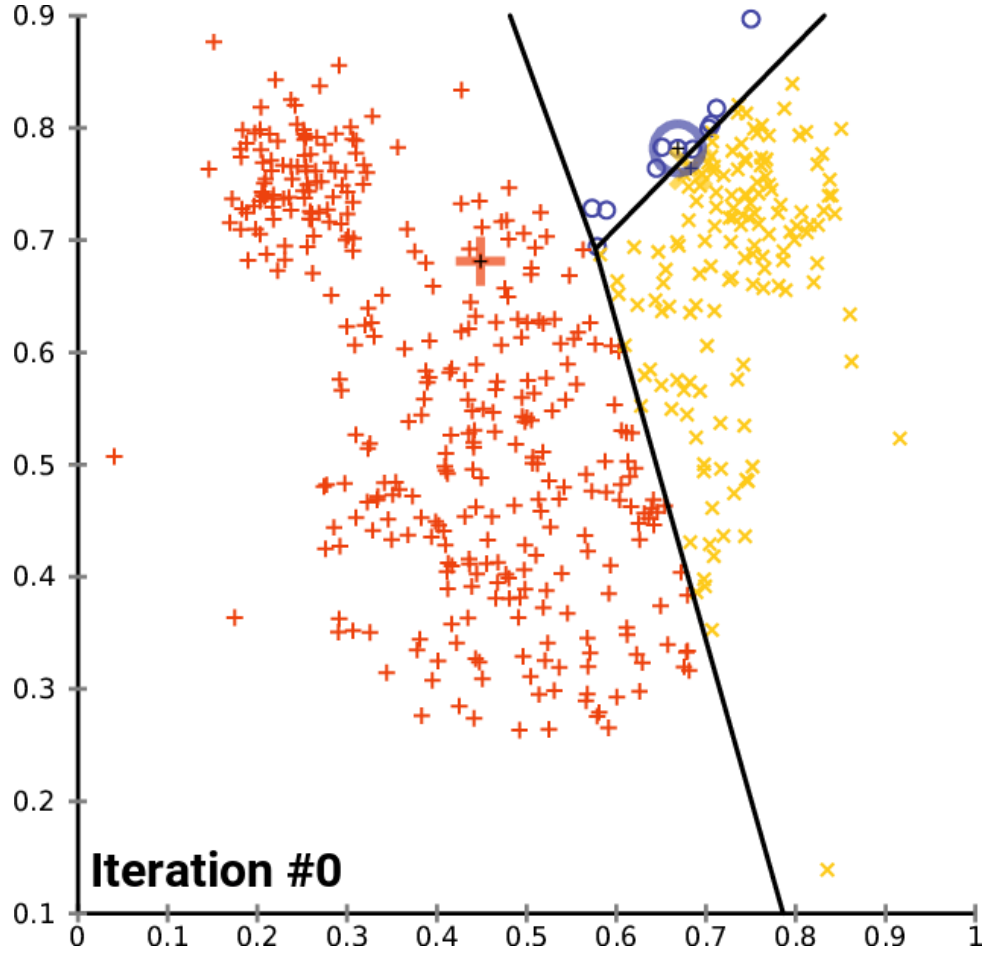
$$\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i' \in C_k} x_{i'j} \iff \sum_{i' \in C_k} x_{i'j} = |C_k| \bar{x}_{kj} \quad (2)$$

□

K-Means Clustering | Beispiel: K = 3



K-Means Clustering | Beispiel: K = 3



Quelle: https://de.wikipedia.org/wiki/K-Means-Algorithmus

K-Means Clustering

Wichtigste bekannte Schwachstellen:

- K-Means Clustering hängt ab von der zufälligen Anfangsinitialisierung der Clusterzugehörigkeiten (Schritt 1) → Findet lokale Minima unserer Optimierungsfunktion, nicht notwendigerweise das globale Minimum.
- Clusteranzahl K muss zu Beginn festgelegt werden. Verschiedene Werte für K können ganz unterschiedliche Cluster liefern.
- Keine Möglichkeit, Ausreißer in den Daten zu erkennen und getrennt zu behandeln.

Empfehlung:

- Starten Sie K-Means mehrfach mit zufälligen initialen Clusterzugehörigkeiten. Wählen Sie das Ergebnis aus, das den kleinsten Wert für die summierte Intra-Cluster Variation aufweist:

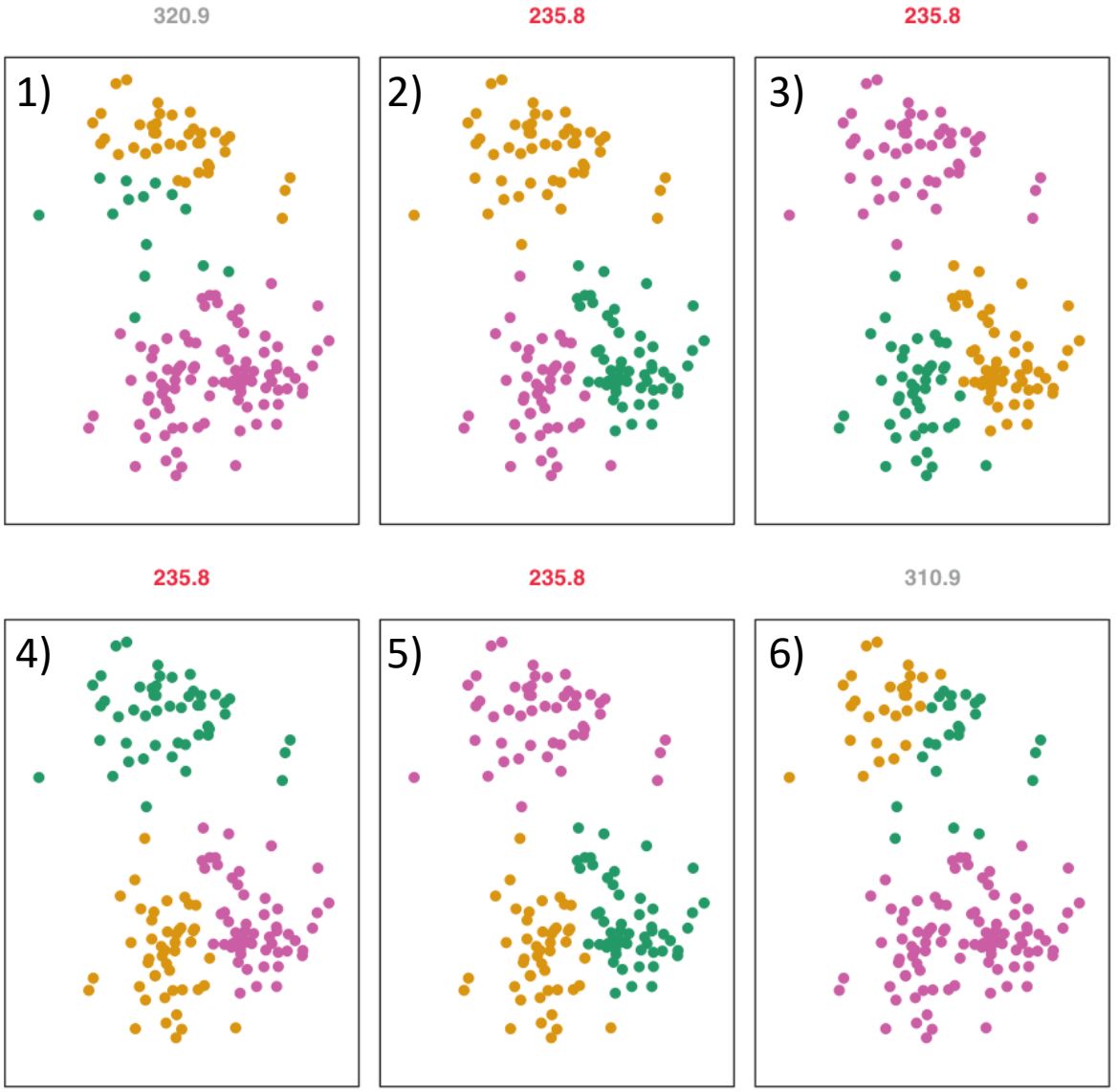
  $\sum_k W(C_k)$ (vgl. Folien zum Optimierungskriterium)

K-Means Clustering | Beispiel (K=3)

Finale Ergebnisse von 6 Läufen von K-Means. Zahlen über den Panels sind die summierten Intra-Cluster Variationen.

Fragen **F**

1. Welches Clustering wählen Sie? 2, 3, 4, oder 5
2. Wieviele lokalen Minima der Optimierungsfunktion sehen Sie hier? 3



Hierarchische Clusteranalyse (HCA)

- einfache (und oft genutzte) Methode zum Finden von Clustern in Daten
- erzeugt eine Menge verschiedener Partitionierungen (=Aufteilung der Daten in Cluster), aus denen eine Partition (Clustering) ausgewählt wird.
- Zentrales Konzept: Aufbau eines Dendrogramms

Dendrogramm: Diagramm, das einen Baum repräsentiert.

Baum: Ungerichteter Graph, in der jedes Paar von Knoten genau durch nur einen Pfad miteinander verbunden ist.

Bäume können aufgebaut werden

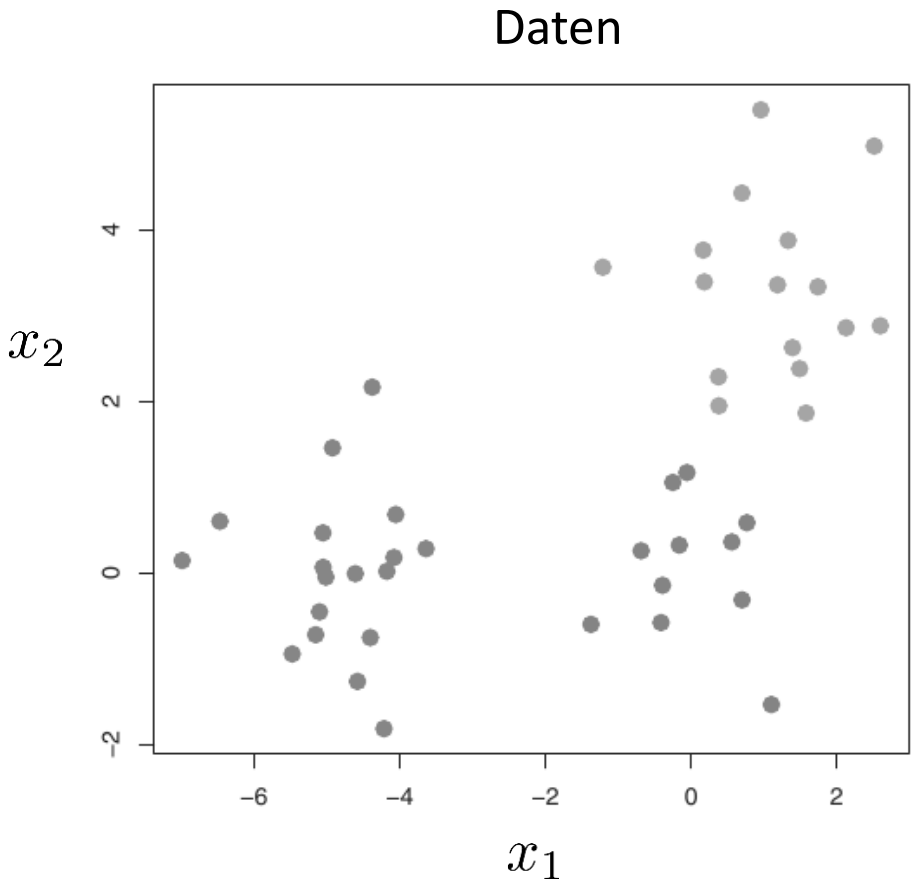
a) von den Blättern hin zur Wurzel
(agglomeratives Clustern, Bottom-Up-Ansatz)

b) von der Wurzel hin zu den Blättern
(divisives Clustern, Top-Down-Ansatz)

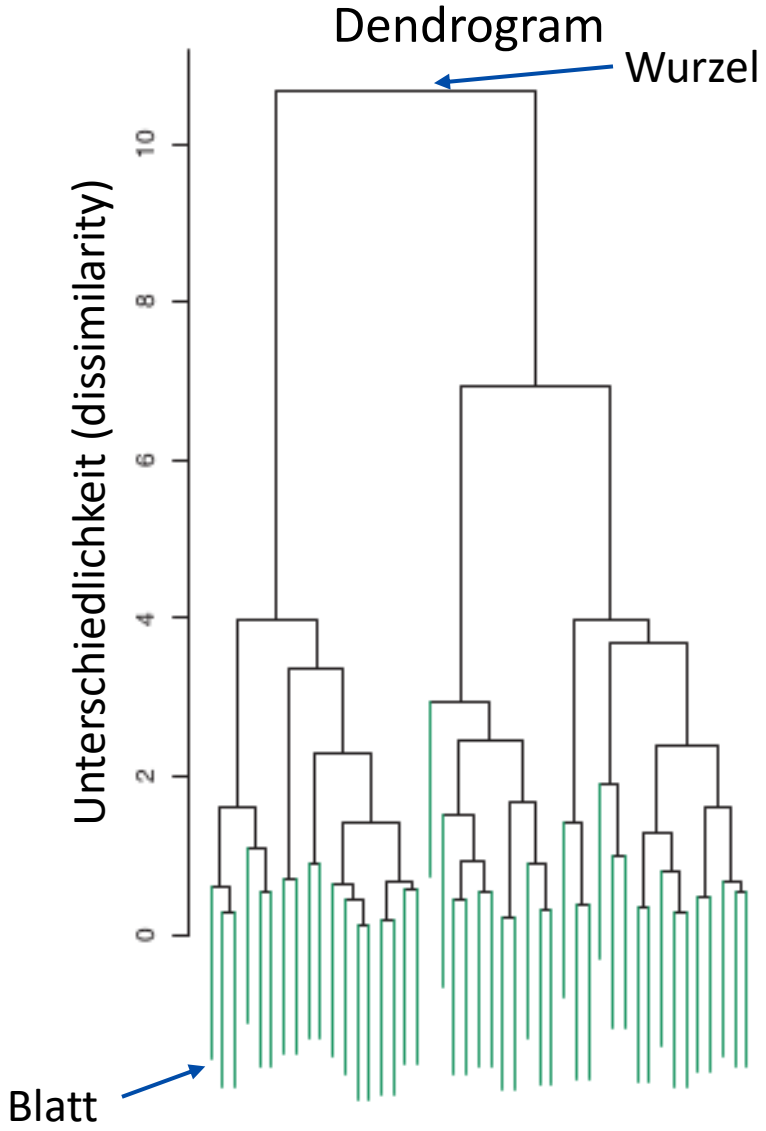
} Vorlesung behandelt diesen Ansatz.

Hierarchisches Clustern | Dendrogramm

Beispiel
(später wird Algorithmus behandelt)

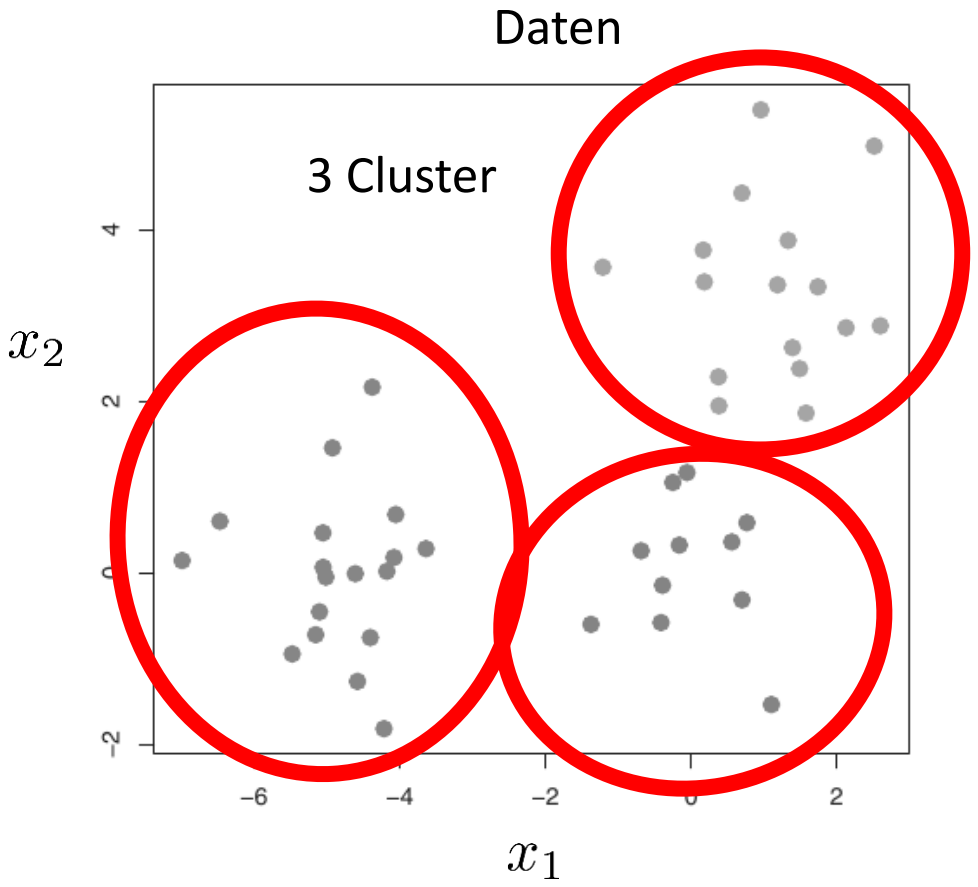


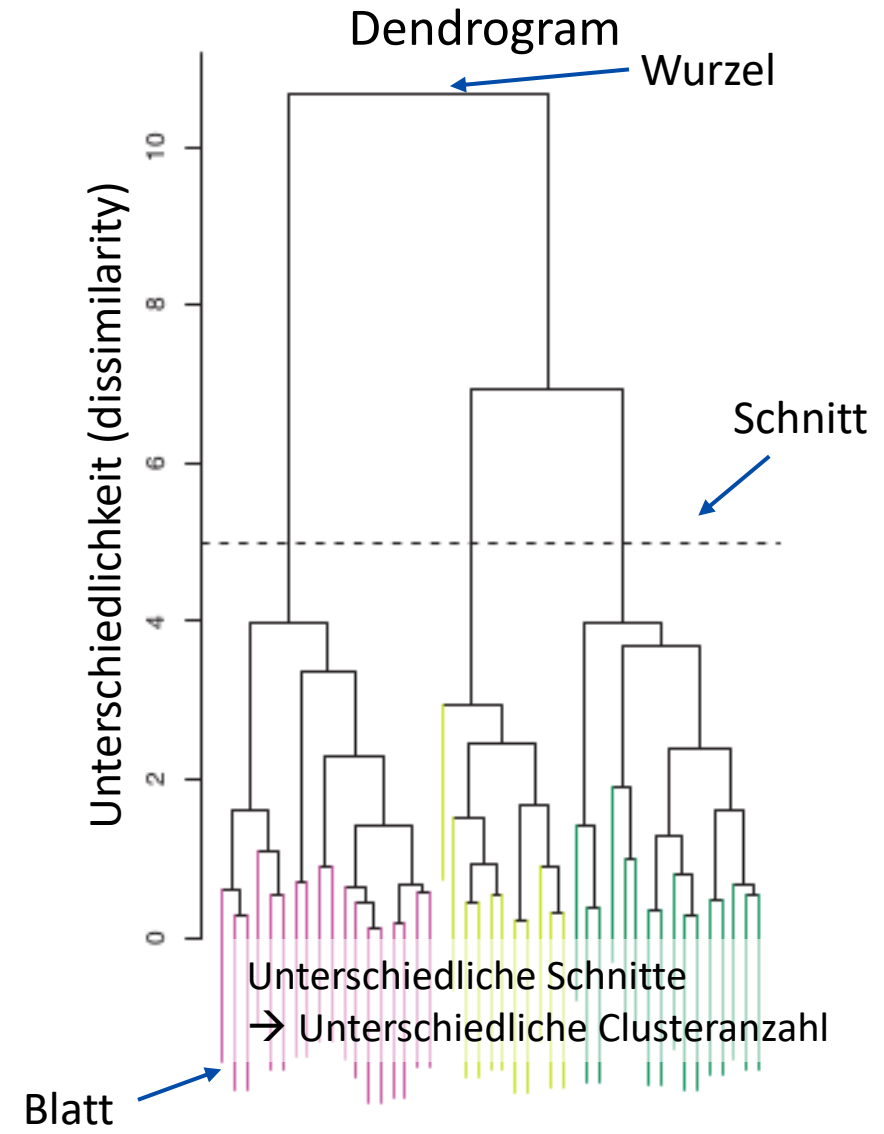
Punkte entsprechen Blättern



Hierarchisches Clustern | Dendrogramm

Beispiel
(später wird Algorithmus behandelt)

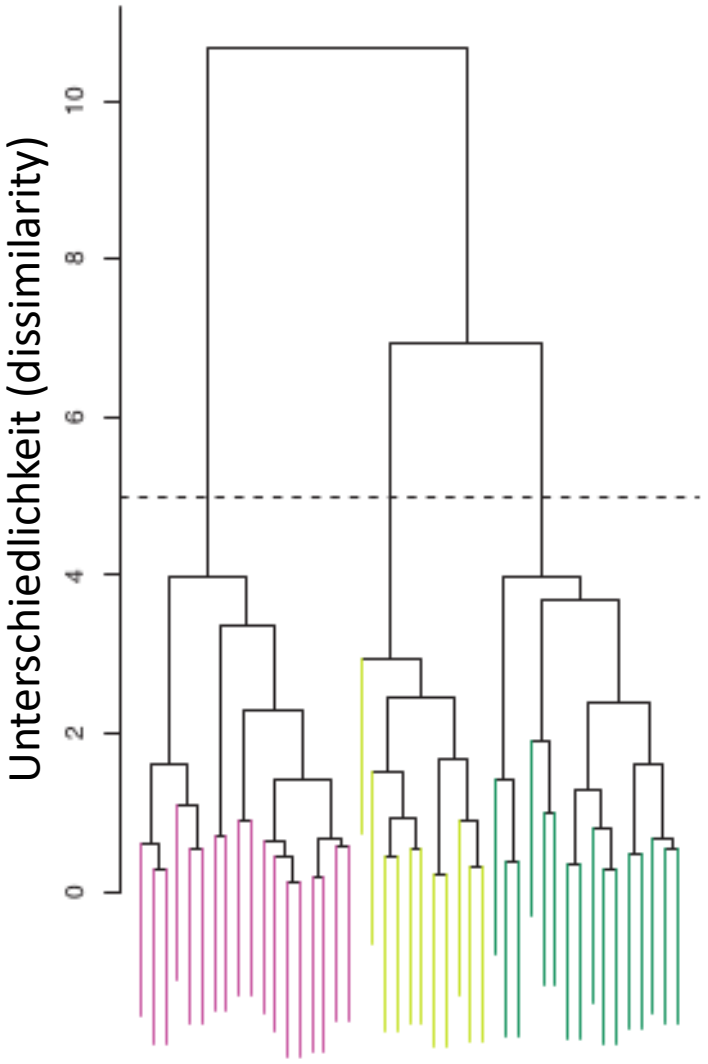




Hierarchisches Clustern | Dendrogramm

- Jeder Datenpunkt entspricht einem Blatt im Dendrogramm.
- Je höher wir im Baum wandern, desto mehr Blätter werden zu Zweigen verbunden.
- Je früher (= weiter unten) Zweige/Blätter verbunden werden, desto ähnlicher sind die entsprechenden Datenpunkte.

- Für jedes Paar von Datenpunkten zeigt die Höhe (y-Achse!) der Verbindung der beiden ihre Unterschiedlichkeit (*dissimilarity*) an.



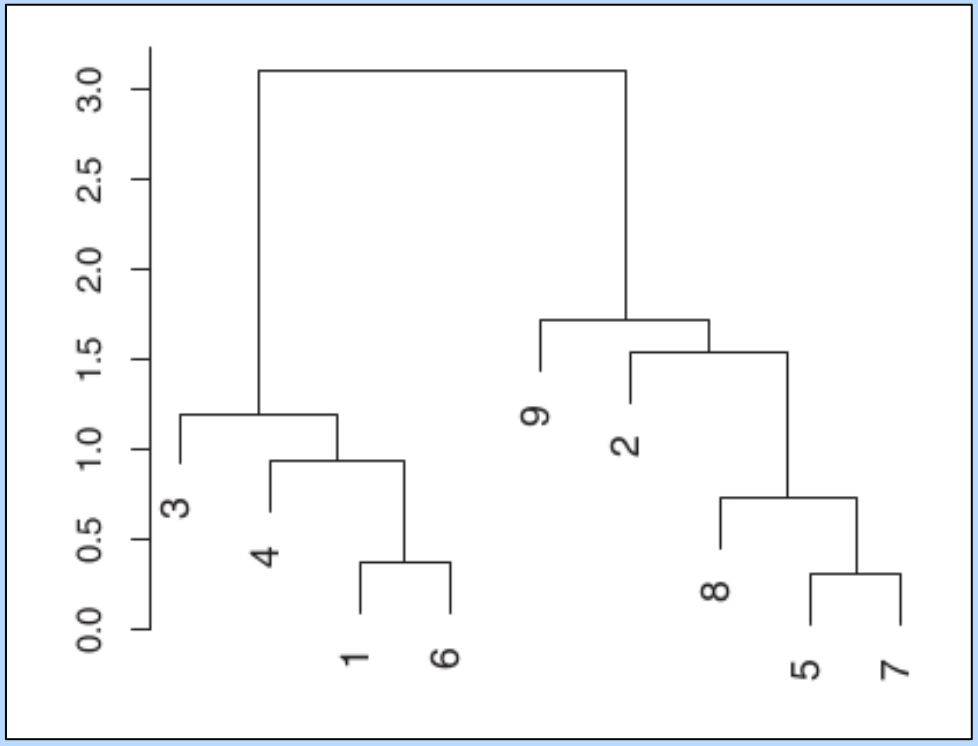
Hierarchisches Clustern | Dendrogramm

F

Frage

Betrachten Sie das dargestellte Dendrogramm.

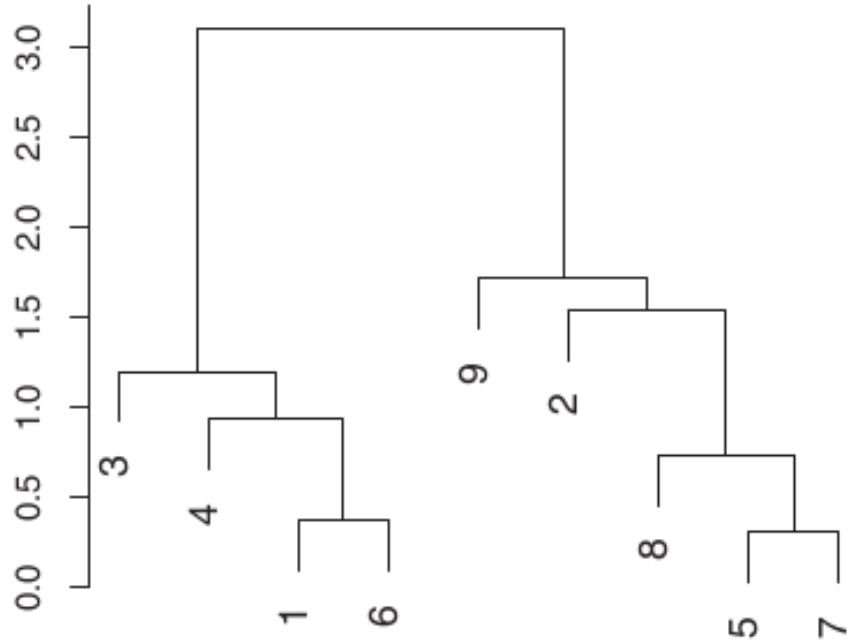
- Sind Datenpunkte 5 und 7 zueinander genauso ähnlich wie Datenpunkte 9 und 2?
  Falls ja, warum?
  Falls nein, warum nicht?



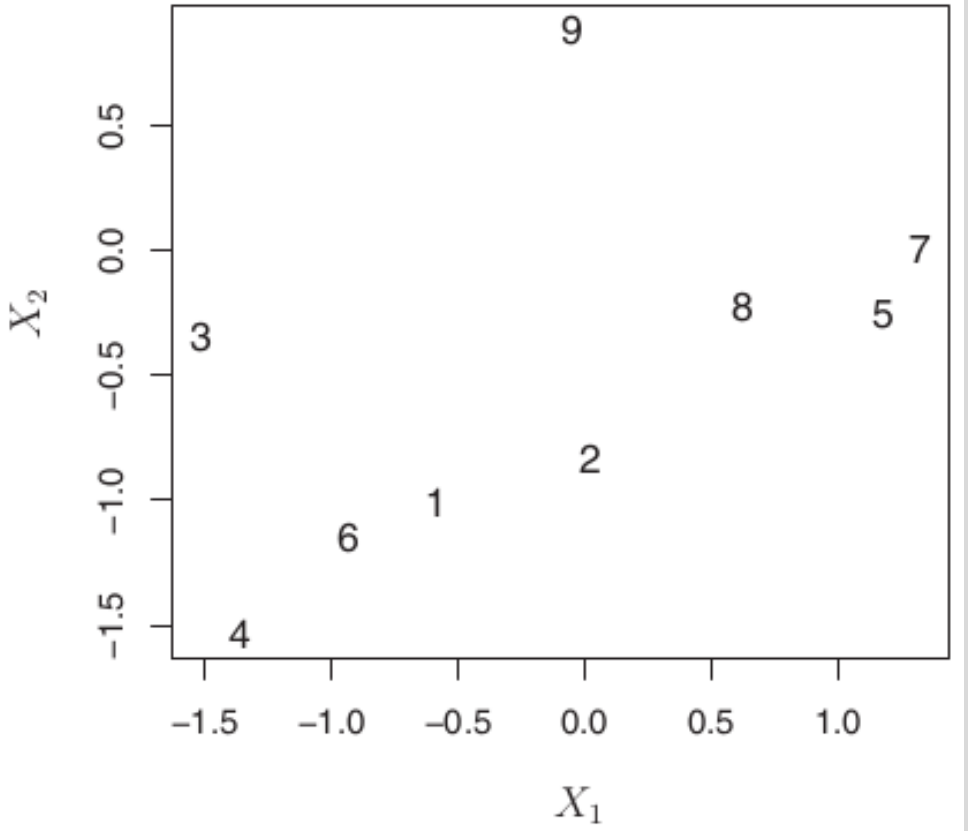
In einem Dendrogramm trägt nur die y-Achse (Höhe) eine Bedeutung, nicht die horizontale Achse!
9 und 2 sind also zueinander deutlich unähnlicher als 5 und 7 (y-Achse betrachten – die x-Achse ist bedeutungslos).

Hierarchisches Clustern | Dendrogramm

Dendrogramm

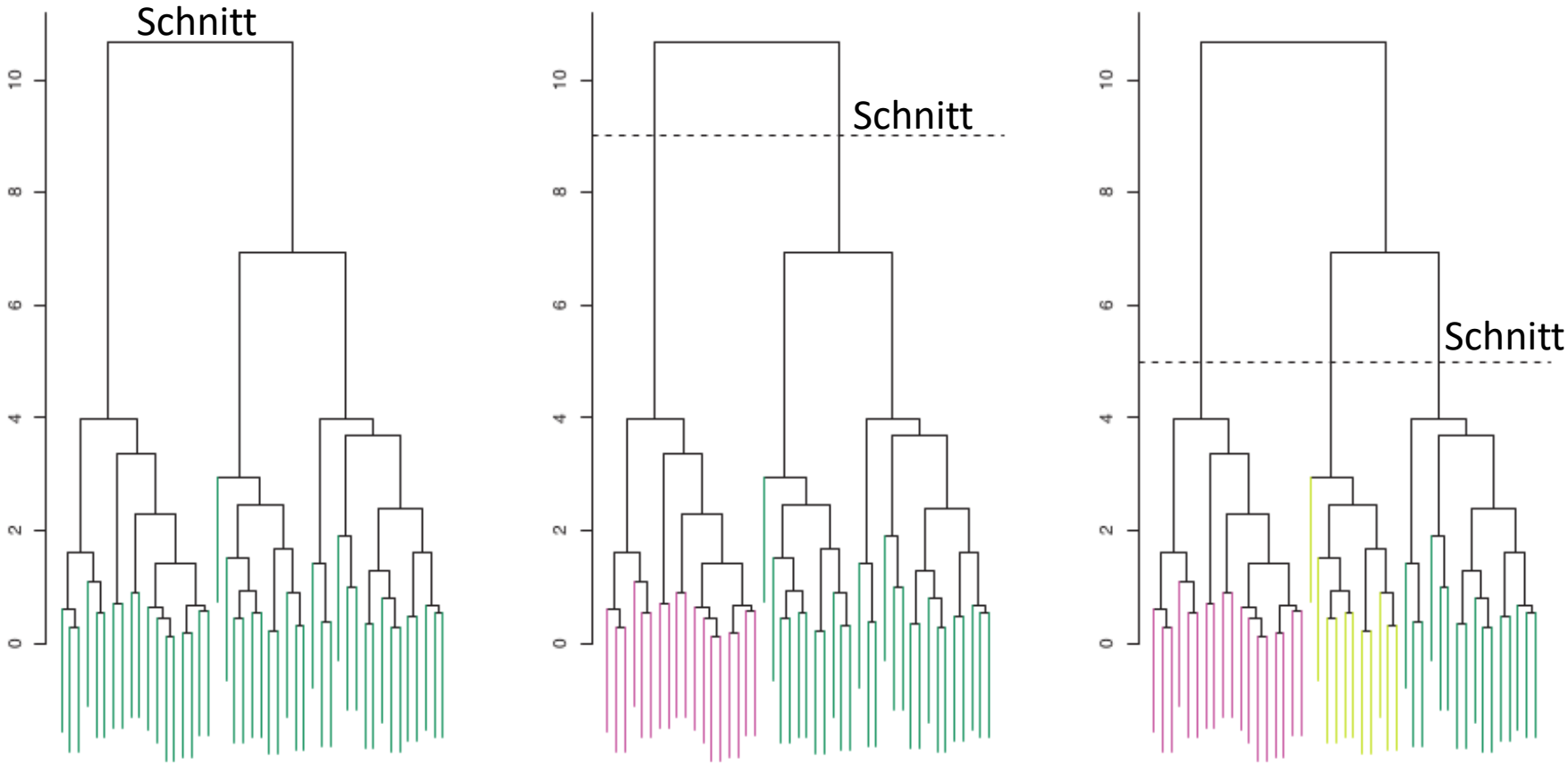


Daten



Hierarchisches Clustern | Dendrogramm

- Schnitte entscheiden, wie viele Cluster die Methode findet (Schnitthöhe ist freier Parameter – so wie K freier Parameter bei K-Means ist)



Hierarchisches Clustern

Algorithmus (agglomeratives hierarchisches Clustern)

1. Betrachte alle *N* Datenpunkte und interpretiere sie als *N* separate Cluster. Wähle ein Maß für die Unterschiedlichkeit (dissimilarity measure) aus.

2. Für i = *N*, *N*-1, … , 2:

   a) Bestimme die paarweisen Unterschiedlichkeiten für alle Cluster.

   b) Fusioniere das Clusterpaar, das sich am ähnlichsten ist, zu einem neuen Cluster. Die Unterschiedlichkeit zwischen den beiden soeben fusionierten Clustern entspricht der Höhe im Dendrogramm, in der der Zusammenschluss stattfindet.

Hierarchisches Clustern

Wir müssen folgende Größen quantifizieren:

1. Unterschiedlichkeit auf Ebene der Datenpunkten
   → **dissimilarity measure** (Unterschiedlichkeitsmaß)

   - hat wichtigen Einfluss auf das Clustering-Ergebnis
   - oft wird euklidische Distanz gewählt

2. Unterschiedlichkeit auf Ebene der Cluster
   → *Linkage* (Verknüpfung)

   - *Linkage* entscheidend für Cluster-Ergebnis
   - Drei bekannte Linkage-Arten: *complete, single, average*

Hierarchisches Clustern | Linkage

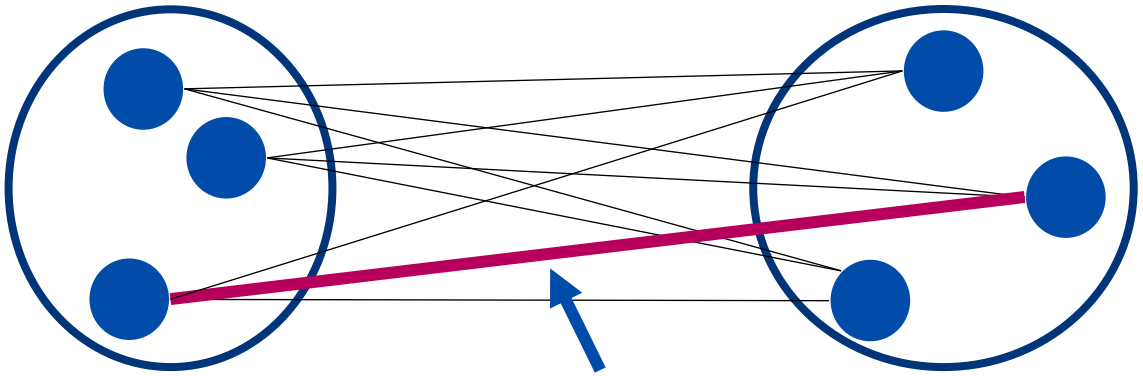
Complete Linkage

= maximale Inter-Cluster-Unterschiedlichkeit

1. Alle paarweisen Unterschiedlichkeiten zwischen Punkten aus dem ersten und Punkten aus dem zweiten Cluster bestimmen.
2. Unterschiedlichkeit zwischen beiden Clustern entspricht der **größten** Unterschiedlichkeit aus Schritt (1).

Beispiel

Unterschiedlichkeitsmaß: euklidische Distanz



Maximale euklidische Distanz entspricht Unterschiedlichkeit beider Cluster

Hierarchisches Clustern | Linkage

Single Linkage

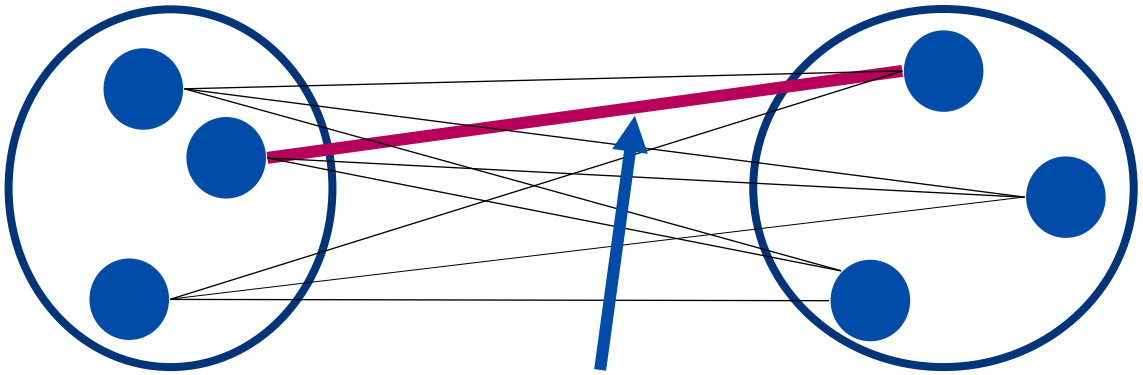
= minimale Inter-Cluster-Unterschiedlichkeit

1. Alle paarweisen Unterschiedlichkeiten zwischen Punkten aus dem ersten und Punkten aus dem zweiten Cluster bestimmen.
2. Unterschiedlichkeit zwischen beiden Clustern entspricht der **kleinsten** Unterschiedlichkeit aus Schritt (1).

- Single Linkage liefert oft (leider) unausgewogene Cluster.

Beispiel

Unterschiedlichkeitsmaß: euklidische Distanz



Minimale euklidische Distanz entspricht Unterschiedlichkeit beider Cluster

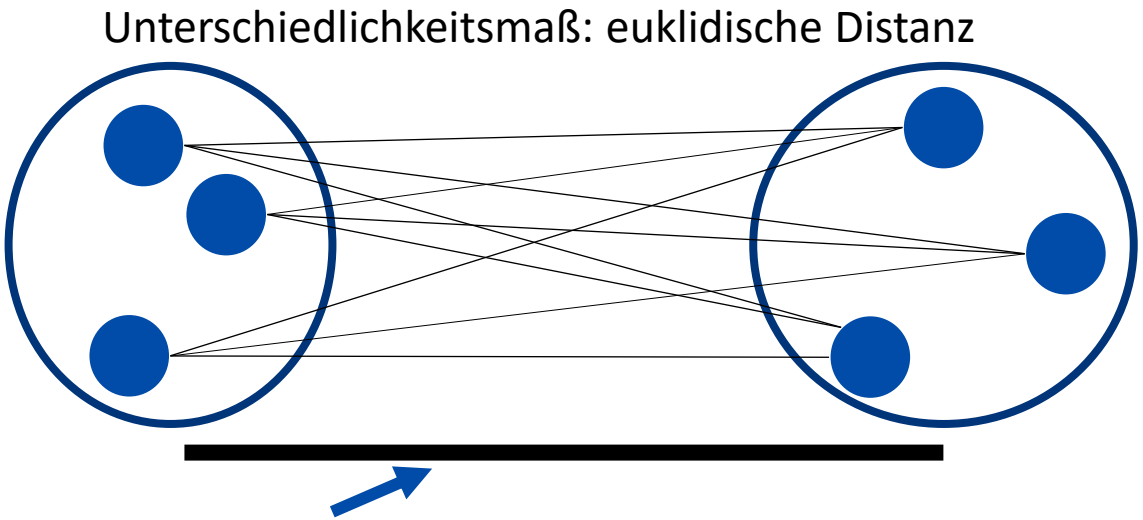
Hierarchisches Clustern | Linkage

Average Linkage

= mittlere Inter-Cluster-Unterschiedlichkeit

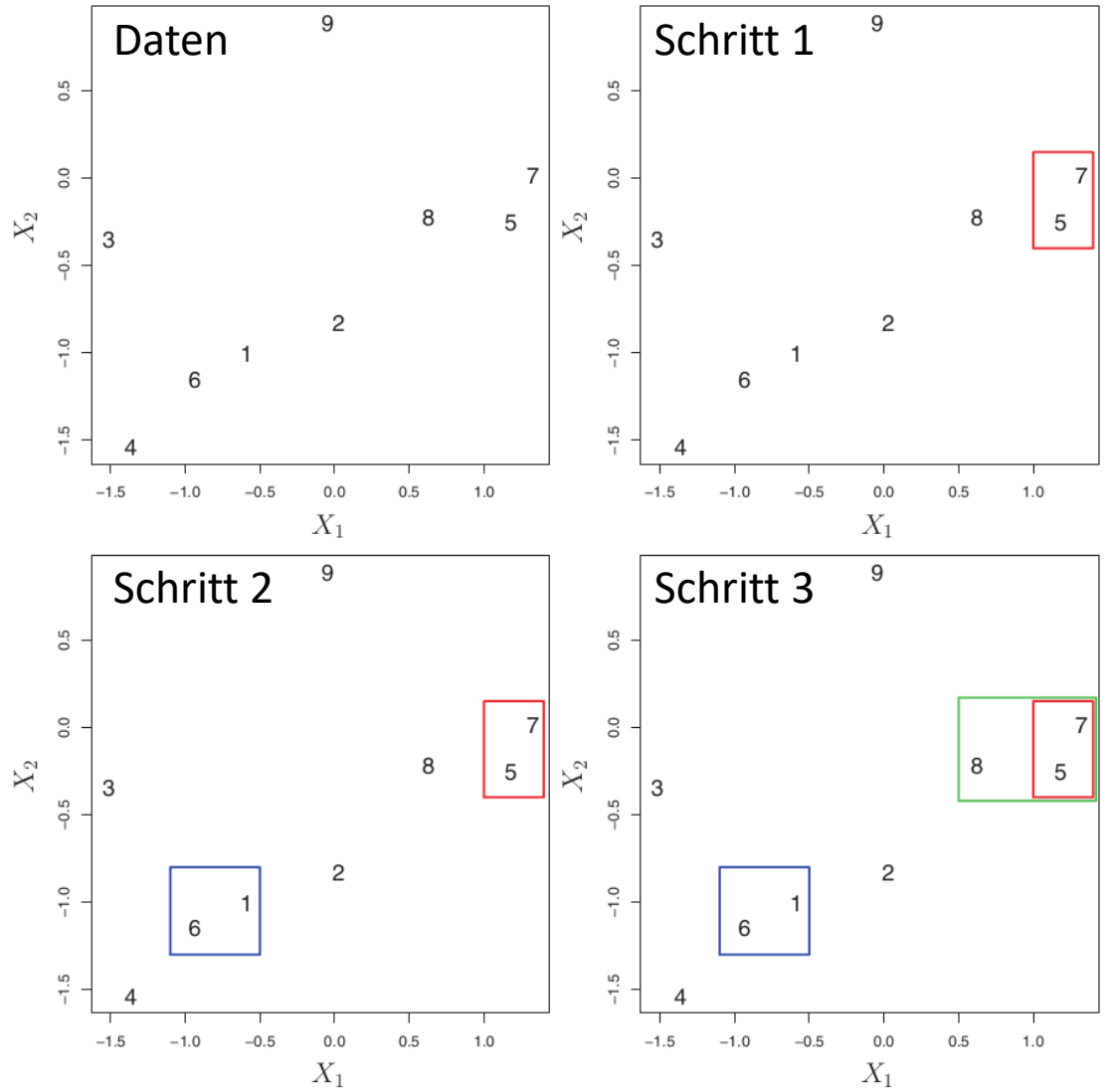
1. Alle paarweisen Unterschiedlichkeiten zwischen Punkten aus dem ersten und Punkten aus dem zweiten Cluster bestimmen.
2. Unterschiedlichkeit zwischen beiden Clustern entspricht dem **Mittelwert** der Unterschiedlichkeiten aus Schritt (1).

Beispiel

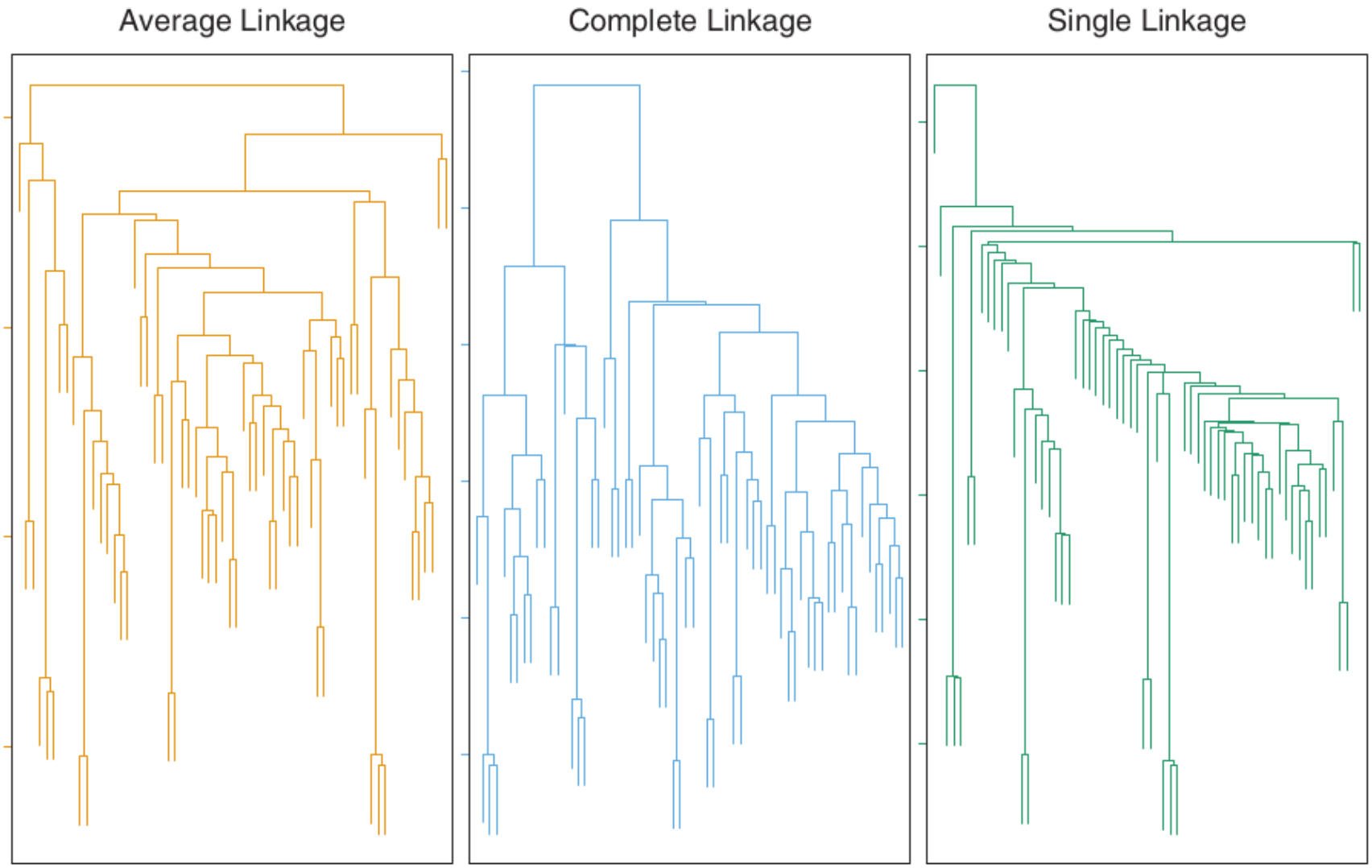


Hierarchisches Clustern | Beispiel

Complete Linkage mit euklidischer Distanz als Unterschiedlichkeitsmaß.



Hierarchisches Clustern



Average und Complete Linkage erzeugen typischerweise ausgewogenere Cluster als Single Linkage.

K-Means vs Hierarchisches Clustern

Clustering Verfahren interpretieren die Daten mit den ihnen impliziten und expliziten Annahmen, wie Cluster zu definieren sind.

Hierarchisches Clustern: Annahme, dass Cluster hierarchisch organisiert sind (d.h. große Cluster setzen sich aus kleineren Clustern zusammen).

K-Means Clustering: Annahme, dass Cluster sphärische Objekte im Feature-Space sind und ungefähr gleich viele Datenpunkte enthalten.

Frage

F

Sei ein Datensatz mit Messungen von 50 Frauen und 50 Männern gegeben, die drei Nationalitäten (Deutsch, Japanisch, Amerikanisch) angehören. Die beste Partitionierung mit 2 Clustern liefert die Geschlechter; die beste Partitionierung mit 3 Clustern liefert die Nationalitäten. Welches Clusterverfahren hat diese Partitionierungen vermutlich erzeugt?

K-Means, weil Cluster nicht hierarchisch organisiert sind. 3 Cluster resultieren nicht aus der Teilung von 2 Cluster.

Clustering-Analysis

Worauf Sie achten sollten:

1. Wahl des Unterschiedlichkeitsmaßes (dissimilarity measure)

Beispiel

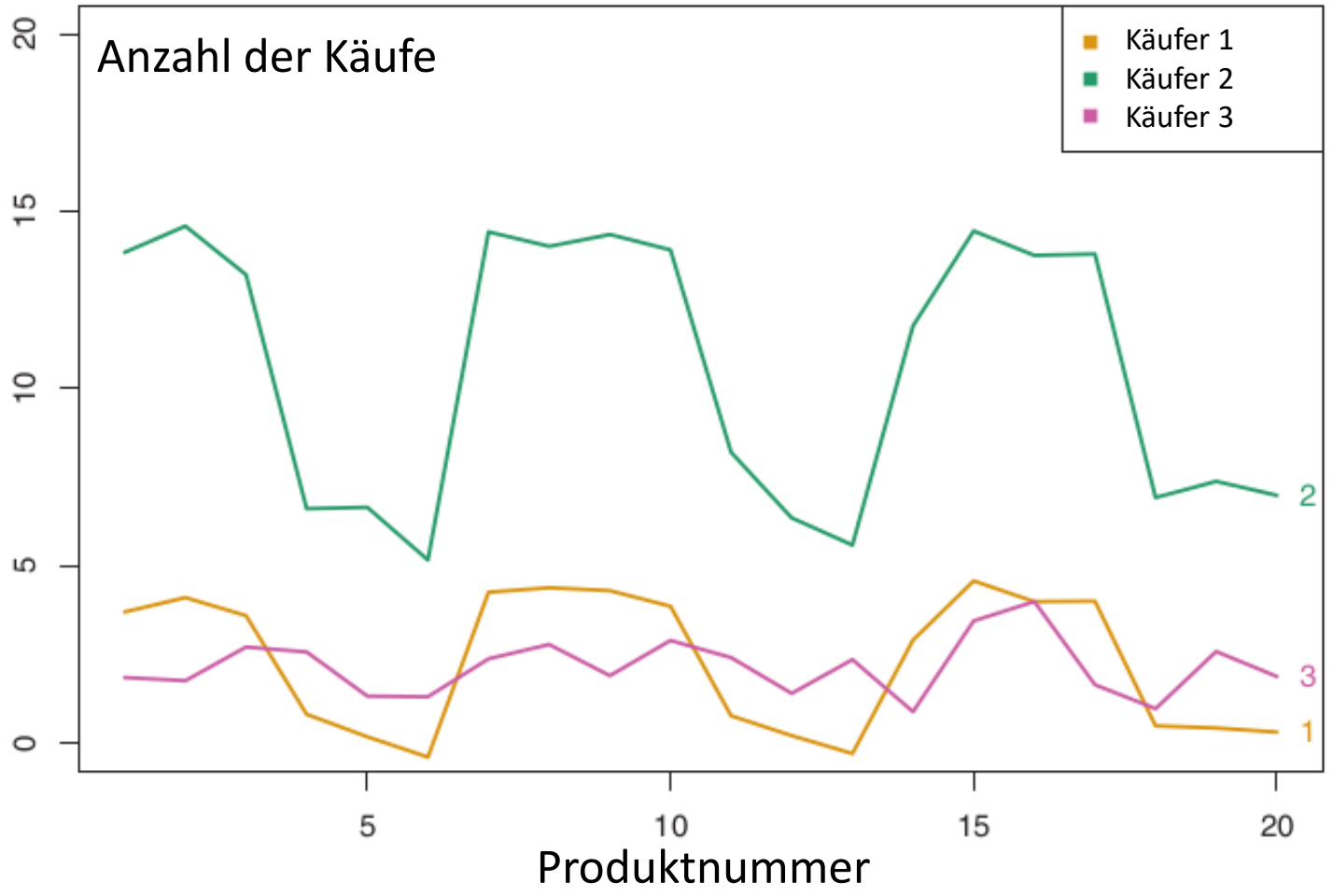
- Online-Händler möchte Kundengruppen durch Clustering erhalten.
- Features eines Kunden sind Anzahl der Käufe für verschiedene Produkte.
- Als *dissimilarity measure* wird die euklidische Distanz benutzt.

Konsequenz

- Feature-Vektoren der Kunden, die keine oder sehr wenige Käufe getätigt haben, liegen nah zueinander im Feature-Raum: Cluster der Wenig- oder Nicht-Käufer wird identifiziert.
- Käufer mit ähnlichem Kaufprofil aber unterschiedlicher Kaufanzahl werden tendenziell nicht im selben Cluster sein (→ dies wäre aber für Kaufempfehlungen wünschenswert).

Clustering Analysis

Beispielhafte (hypothetische) Feature-Vektoren von Käufern



Euklidische Distanz: Käufer 1 und 3 sind im selben Cluster, weil beide wenig kaufen
Korrelation als Distanzmaß: Käufer 1 und 2 sind im selben Cluster, weil beide ein ähnliches Kaufverhalten zeigen.

Clustering-Analysis

Worauf Sie achten sollten:

2. Vorverarbeitung der Daten

- Untersuchen Sie, welche Auswirkungen eine Skalierung der Features in Kombination mit der Wahl Ihres Unterschiedlichkeitsmaß (dissimilarity measure) auf Cluster-Ergebnisse hat.

3. Wahl der Clusteranzahl (oder anderer freier Parameter)

- Untersuchen Sie, welche Auswirkungen Ihre Wahl einer Clusteranzahl auf die Ergebnisse Ihrer Cluster-Analyse hat.

4. Interpretation Ihrer Ergebnisse

- Sehen Sie eine Cluster-Analyse als Startpunkt einer Untersuchung der Daten, nicht als Endpunkt einer Analyse.