

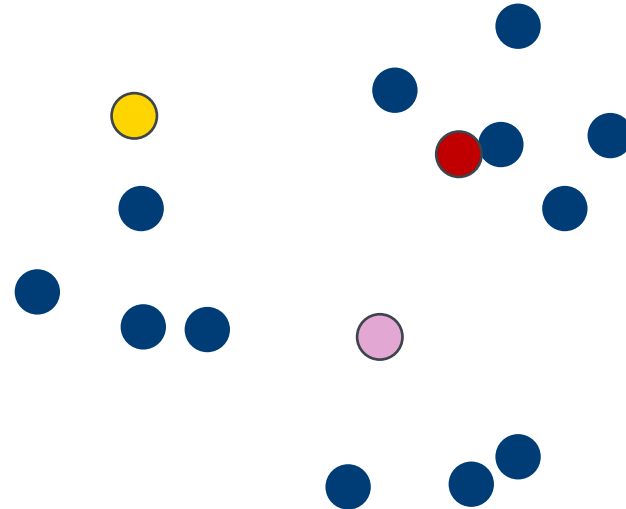
Clusteranalyse – Verfahren

Beispiel k-Means

Beispiel k-Means

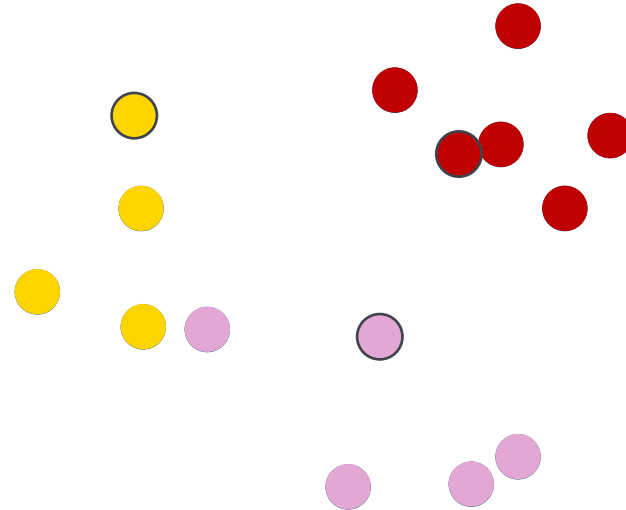
k-Means Algorithmus – Initialisierung

- Wähle gewünschte Anzahl von Clustern
- Wähle zufällig entsprechend viele Clustermittelpunkte



K-Means Algorithmus – 1. Schritt

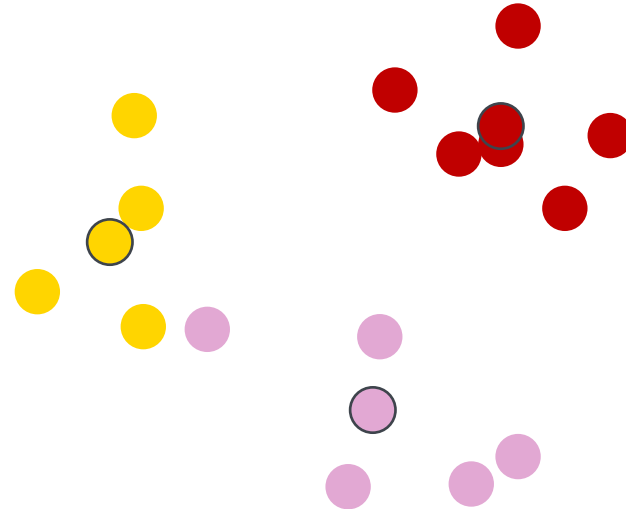
- Ordne jeden Punkt dem nächstliegenden Clustermittelpunkt zu



K-Means Algorithmus – 2. Schritt

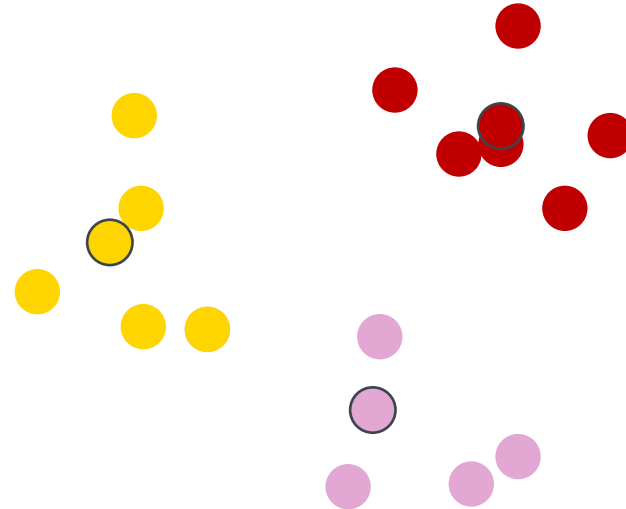
- Finde für jedes Cluster den neuen Mittelpunkt

Schritt 1 und 2
wiederholen, bis
sich nichts mehr
ändert



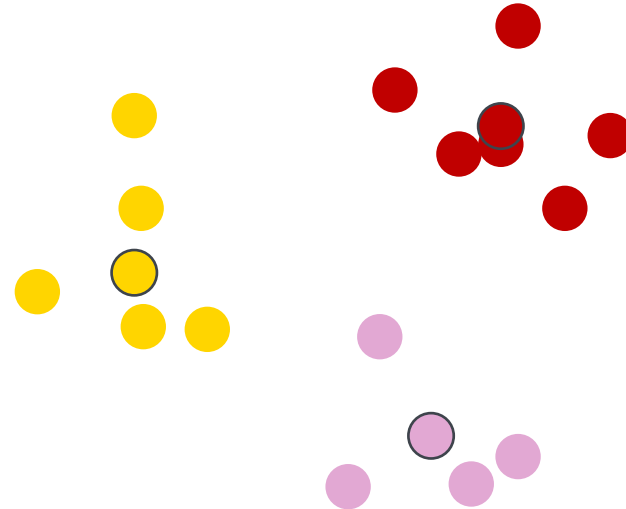
K-Means Algorithmus – 1. Schritt

- Wiederhole Schritt 1:
- Ordne jeden Punkt dem nächstliegenden Clustermittelpunkt zu



K-Means Algorithmus – 2. Schritt

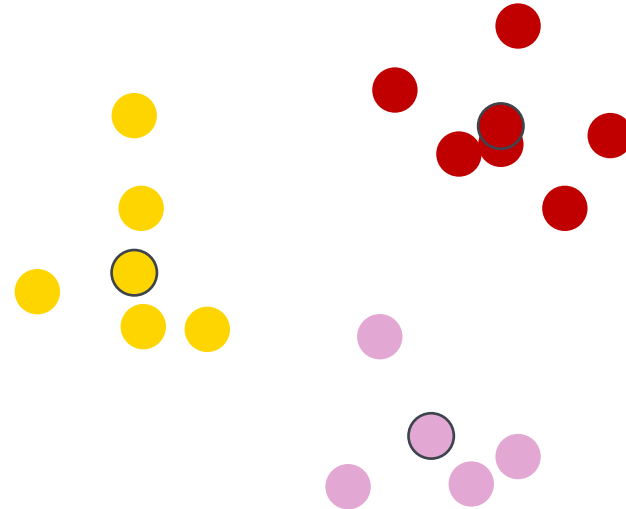
- Wiederhole Schritt 2:
- Finde für jedes Cluster den neuen Mittelpunkt



K-Means Algorithmus – Ende

- Die Zuordnung ändert sich nicht mehr
- Also auch nicht die Mittelpunkte

Der Algorithmus
konvergiert -
fertig!



**Beim Clustern mit k-Means
werden zur Initialisierung
k Clustermittelpunkte („Means“)
zufällig gewählt.**

**Die Anzahl der Cluster (k) muss
also im Voraus festgelegt werden.**

Nach der Initialisierung werden folgende Schritte wiederholt, bis der Algorithmus konvergiert:

Ordne jeden Datenpunkt dem nächstgelegenen Clustermittelpunkt zu.*

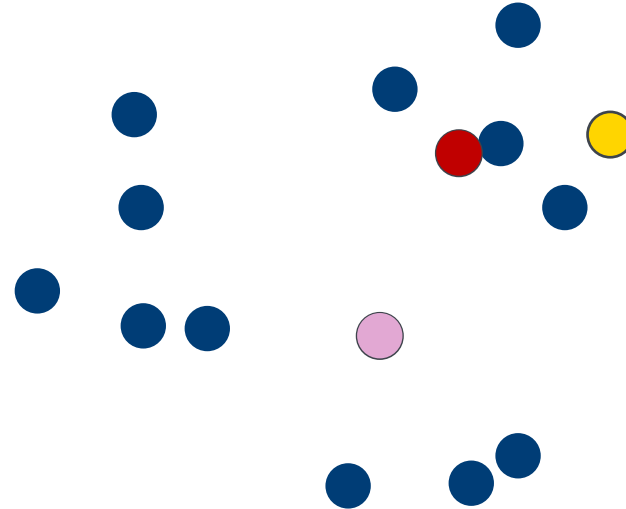
Berechne für jedes Cluster seinen neuen Mittelpunkt.**

*beim klassischen k-Means gemäß der euklidischen Distanz

**beim klassischen k-Means als arithmetisches Mittel aller Punkte im Cluster

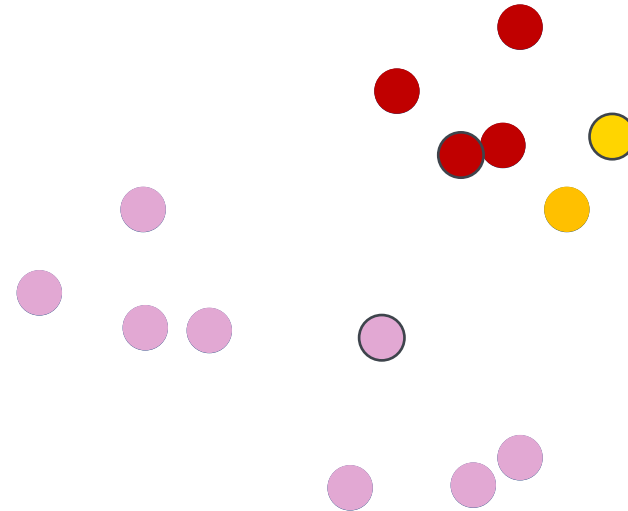
K-Means Algorithmus – Initialisierung

- Es gibt auch ungünstige Initialisierungen



K-Means Algorithmus – 1. Schritt

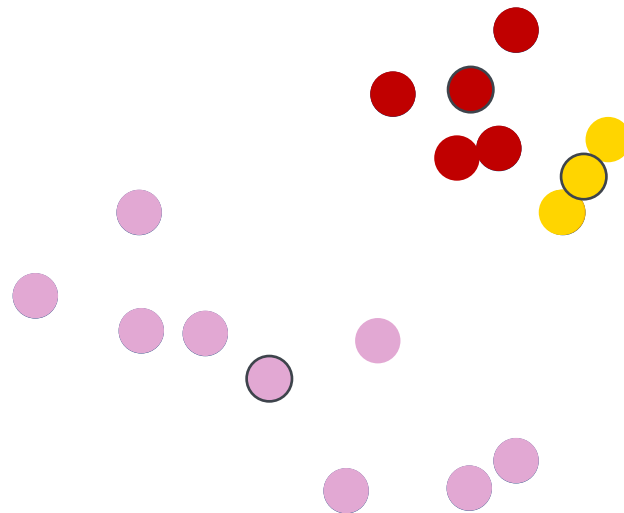
- Ordne jeden Punkt dem nächstliegenden Clustermittelpunkt zu



K-Means Algorithmus – 1. Schritt

- Berechne die neuen Mittelpunkte

Der Algorithmus
konvergiert –
allerdings nicht mit
dem gewünschten
Ergebnis!

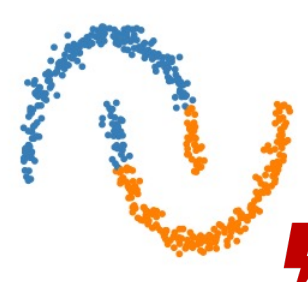
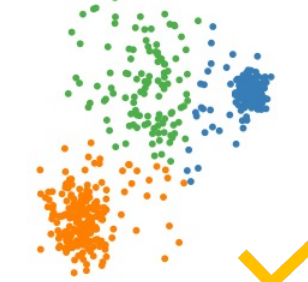
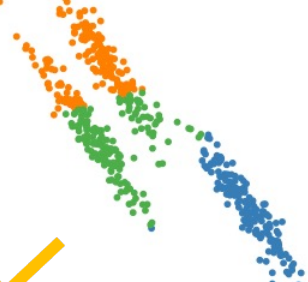
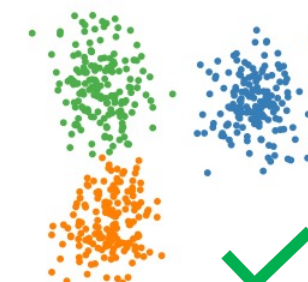
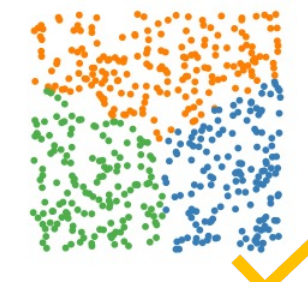


Die Initialisierung hat bei k-Means einen erheblichen Effekt und kann das Ergebnis positiv wie negativ beeinflussen.

Es gibt verschiedene Methoden, möglichst „gute“ Initialisierungen zu finden.

Außerdem wird oft in mehreren Durchläufen mit verschiedenen Initialisierungen geclustert und dann das „beste“ Ergebnis ausgewählt.

Wann funktioniert k-Means gut?

Distanz kein gutes Kriterium				Dichte spielt eine Rolle
Skalierung problematisch				Anzahl der Cluster nicht sinnvoll

**Wie gut k-Means geeignet ist, hängt
von der Verteilung der Daten ab.**

**Es gibt Fälle, in denen andere
Cluster-Algorithmen besser
geeignet sind.**

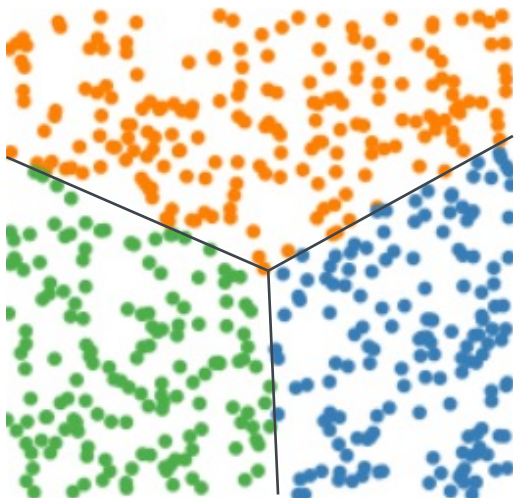
Bewertung k-Means

- Der k-Means Algorithmus ist ein sehr einfacher Algorithmus zum Auffinden von Clustern
- Er basiert auf (euklidischen) Distanzen zwischen Datenpunkten
- Er ist daher nicht geeignet
 - in Fällen, wo Distanzen keine Rolle für die Zuordnung zu Clustern spielen
- Er ist nur bedingt geeignet
 - bei ungünstig skalierte Daten
 - bei Daten, die unwichtige Dimensionen enthalten
- Verschiedene Initialisierungen können unterschiedliche Ergebnisse liefern. Um ein „bestes“ Ergebnis auszuwählen, braucht man ein Evaluationsmaß

Später
mehr!

Ebenfalls
später mehr!

Partitionierende Clusterverfahren



- k-Means ist ein partitionierendes Clusterverfahren
- Es unterteilt den Raum in nicht-überlappende Regionen (bestimmt durch Cluster-Mittelpunkte)
- Jeder Datenpunkt gehört zu genau einem Cluster
- Ebenso ggf. für neue Datenpunkte*

*Achtung, mit neuen Datenpunkten evtl. andere Cluster, falls man eine neue Analyse durchführt

**Partitionierende Clusterverfahren
teilen den Raum in nicht-
überlappende Regionen auf.**

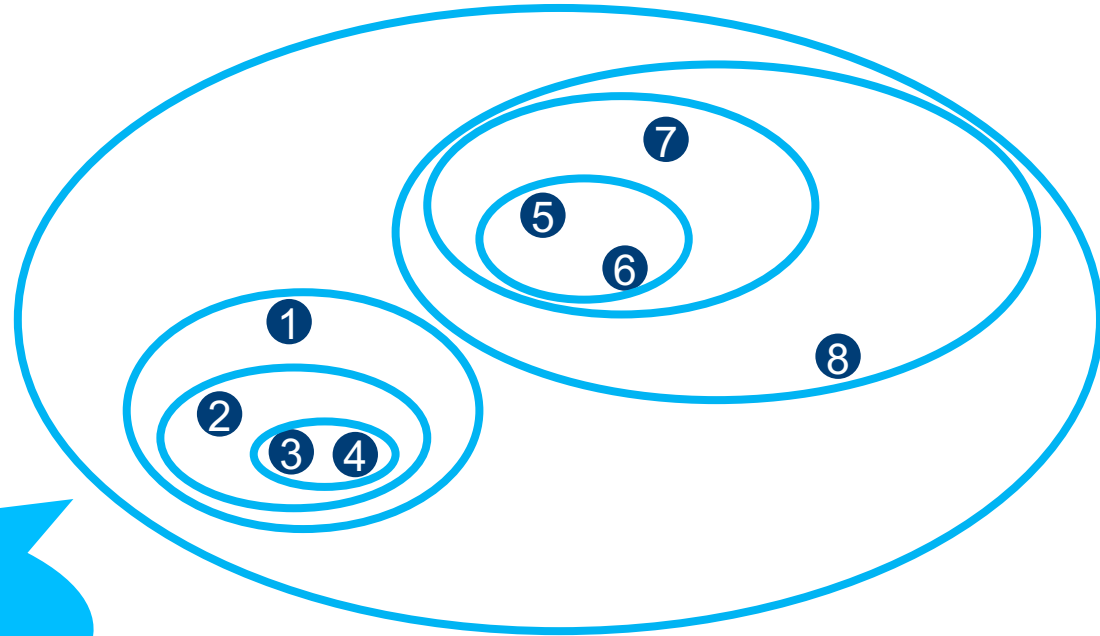
**Die Zuordnung jedes Datenpunkts
zu einem Cluster ist eindeutig.**

Andere Verfahren: Hierarchische Clusteranalyse

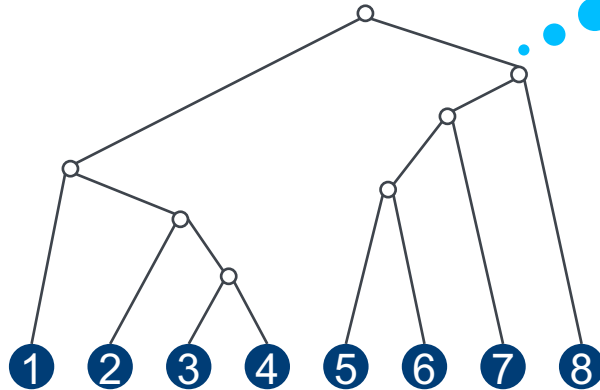
Hierarchische Clusterverfahren

- Starte mit einem Cluster pro Datenpunkt
- Wiederhole laufend: Verbinde die zwei ähnlichsten Cluster zu einem
- Hier: ähnlichste Cluster = Cluster, die zwei Punkte mit minimaler euklidischer Distanz enthalten

Hierarchie von ineinander verschachtelten Clustern

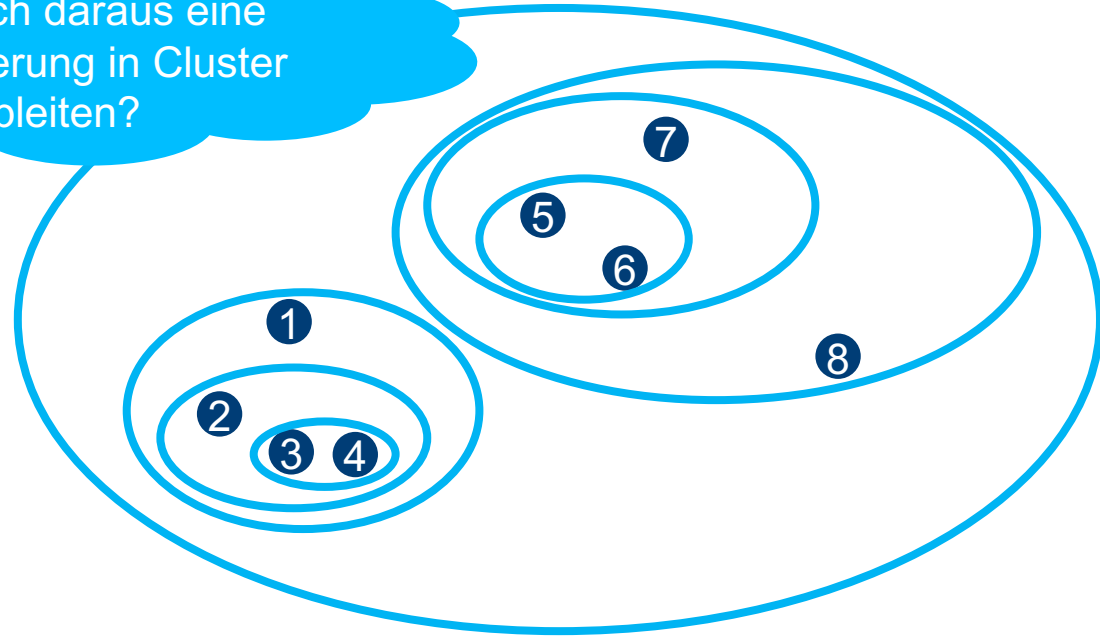


Dendrogramm zur Darstellung der Hierarchischen Clusteranalyse

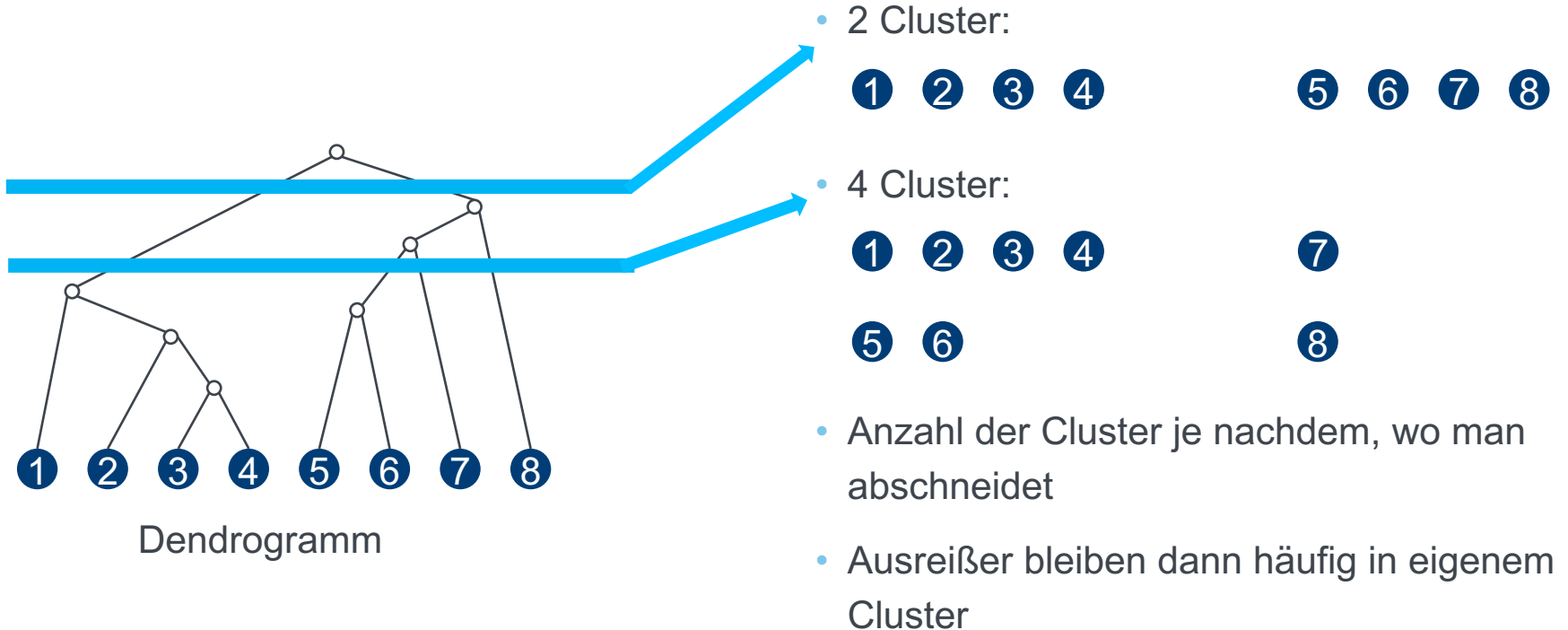


Dendrogramm

Lässt sich daraus eine
Partitionierung in Cluster
ableiten?



Partitionierung beim hierarchischen Clustern



**Hierarchische Clusterverfahren
ergeben eine hierarchische Struktur
von ineinander geschachtelten
Clustern.**

**Datenpunkte können dadurch
mehreren Clustern angehören.**

**Die Ergebnisse der hierarchischem
Clusteranalyse können als
Baumstruktur in einem
Dendrogramm dargestellt werden.**

**Falls eine eindeutige Partitionierung
gewünscht ist, kann im
Dendrogramm nachträglich an einer
bestimmten Stelle „abgeschnitten“
werden.**

Agglomerativ vs. divisiv

- Das vorgestellte Beispiel folgte einem agglomerativen Ansatz
 - Die Struktur wird “von unten” aufgebaut
 - D.h. man beginnt mit einem Punkt pro Cluster
 - Anschließend werden solange je zwei Cluster miteinander verschmolzen, bis es nur noch ein großes Cluster gibt
- Divisive Ansätze gehen umgekehrt vor
 - Man beginnt mit allen Punkten in einem einzigen Cluster
 - Anschließend wird solange jedes Cluster in zwei möglichst unähnliche Cluster aufgeteilt, bis es pro Punkt ein Cluster gibt
 - D.h. die Struktur wird „von oben“ aufgebaut

Dr. Antje Schweitzer

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung



Universität Stuttgart

Institut für Maschinelle Sprachverarbeitung
Institut für Software Engineering



Industrie- und Handelskammer
Reutlingen

Reutlingen | Tübingen | Zollernalb



Region Stuttgart



Industrie- und Handelskammer
Karlsruhe



LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Lizenzbestimmungen

“Clusteranalyse – Verfahren” von Antje Schweitzer, KI B³ / Uni Stuttgart

Das Werk - mit Ausnahme der folgenden Elemente:

- Logos der Verbundpartner und des Förderprogramms
- im Quellenverzeichnis aufgeführte Medien

ist lizenziert unter:

 **CC BY 4.0** (<https://creativecommons.org/licenses/by/4.0/deed.de>)

(Namensnennung 4.0 International)

Quellenverzeichnis

Titelfoto: [Kelly Sikkema](https://unsplash.com/de/@kellysikkema) (<https://unsplash.com/de/@kellysikkema>), Junge im grauen T-Shirt mit Rundhalsausschnitt spielt LEGO-Steine mit weißem Handbuchbuch, auf [Unsplash](https://unsplash.com/de/fotos/junge-im-grauen-t-shirt-mit-rundhalsausschnitt-spielt-lego-steine-mit-weissem-handbuchbuch-Z9AU36chmQI) (<https://unsplash.com/de/fotos/junge-im-grauen-t-shirt-mit-rundhalsausschnitt-spielt-lego-steine-mit-weissem-handbuchbuch-Z9AU36chmQI>), lizenziert unter [Unsplash-Lizenz](https://unsplash.com/license) (<https://unsplash.com/license>). Bildausschnitt verändert.