**8 Clustering**

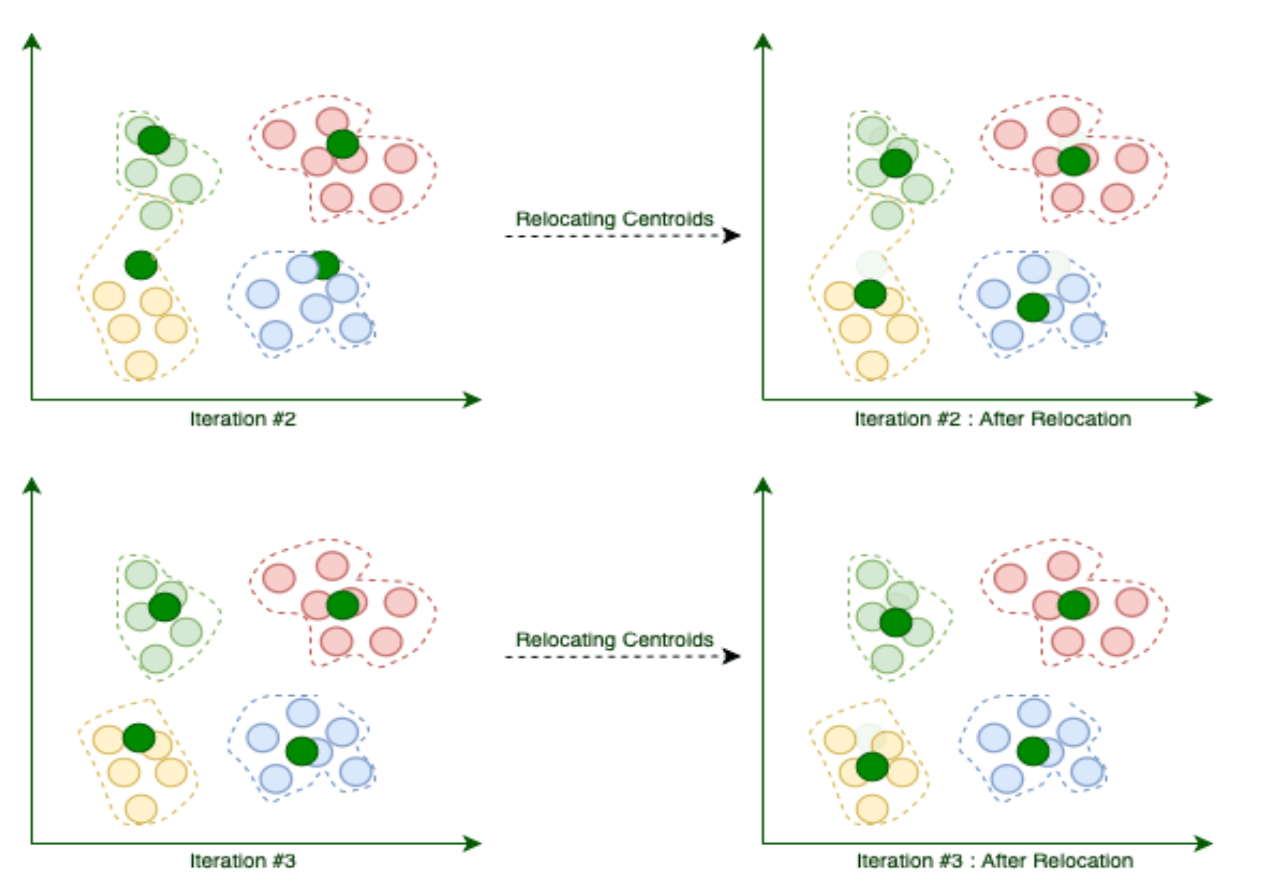
1. K-means clustering

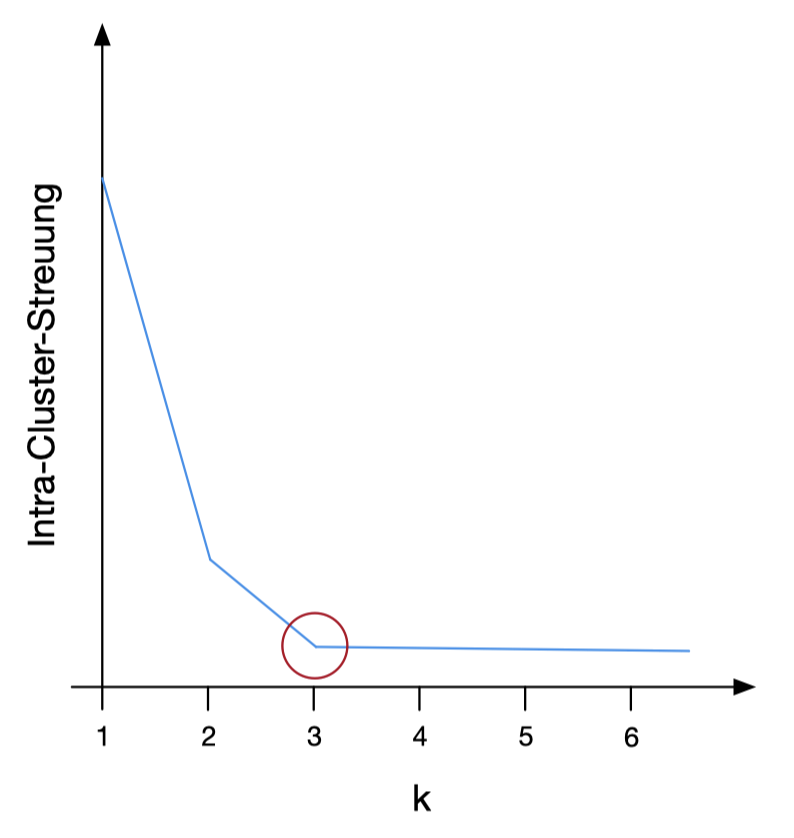
a). k Stichproben als Centroids unseres Clusters im Datensatz zufällig auszuwählen.

b). den Abstand zwischen den anderen Punkten berechnen und weisen die Punkte dem nächsten Centroids zu.

c). die Iteration starten, finden die neuen Centroids dieser Kategorien, indem der Gruppenmittelwert berechnet wird, den Abstand zwischen allen Punkten neu ausrechnen und weisen die Punkte dem nächsten Centroids wieder zu.

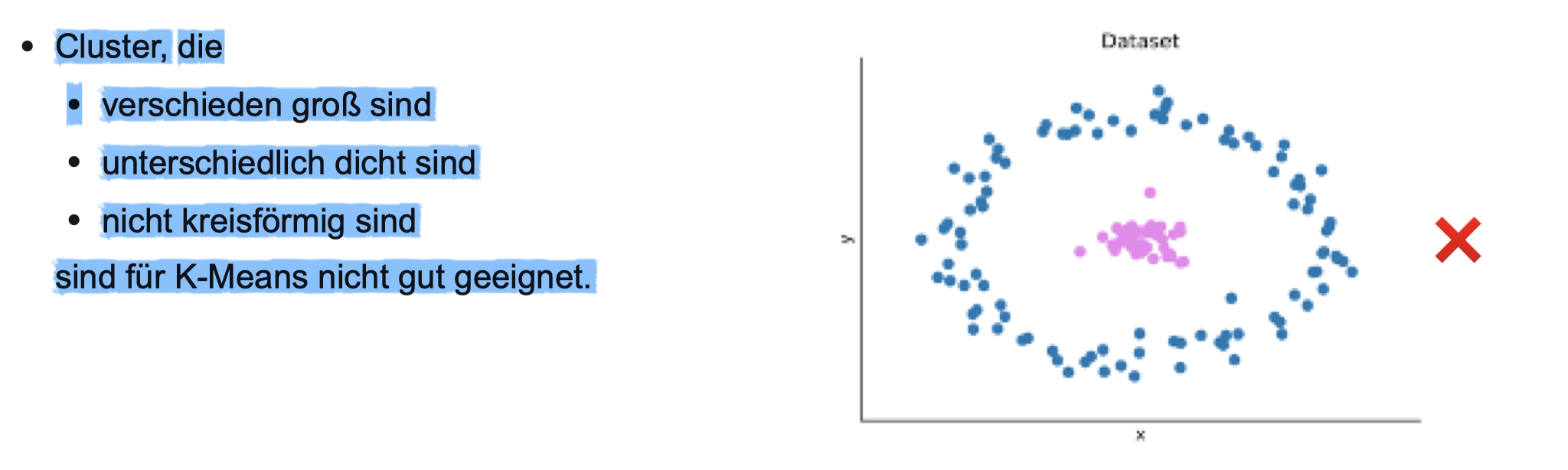
d). Wiederholen die Iteration, bis das Ergebnis eines Clusters mit dem des vorherigen übereinstimmt.



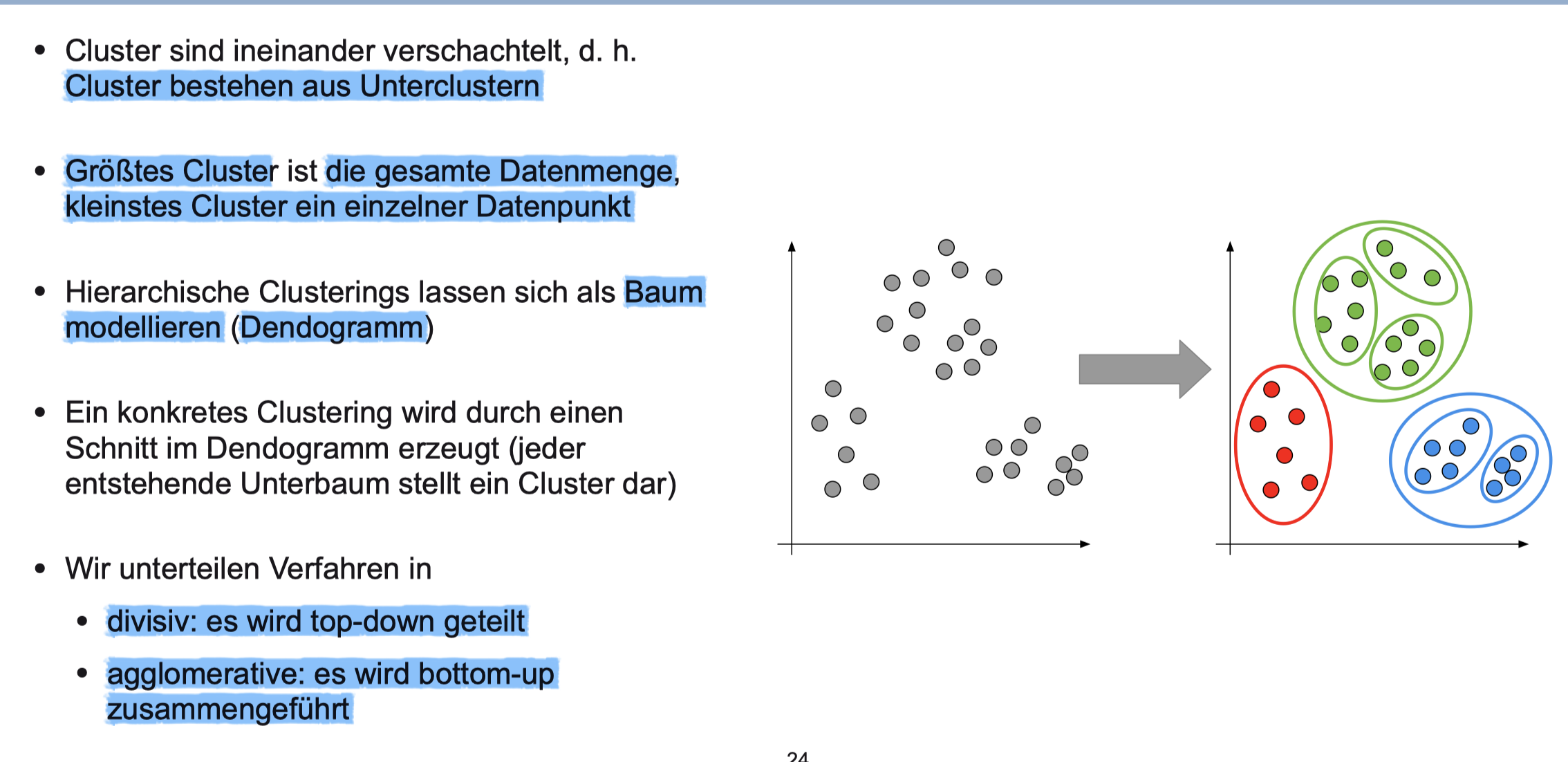


Elbow method, optimal K ausrechnen (SSE: Sum of Squard Errors)

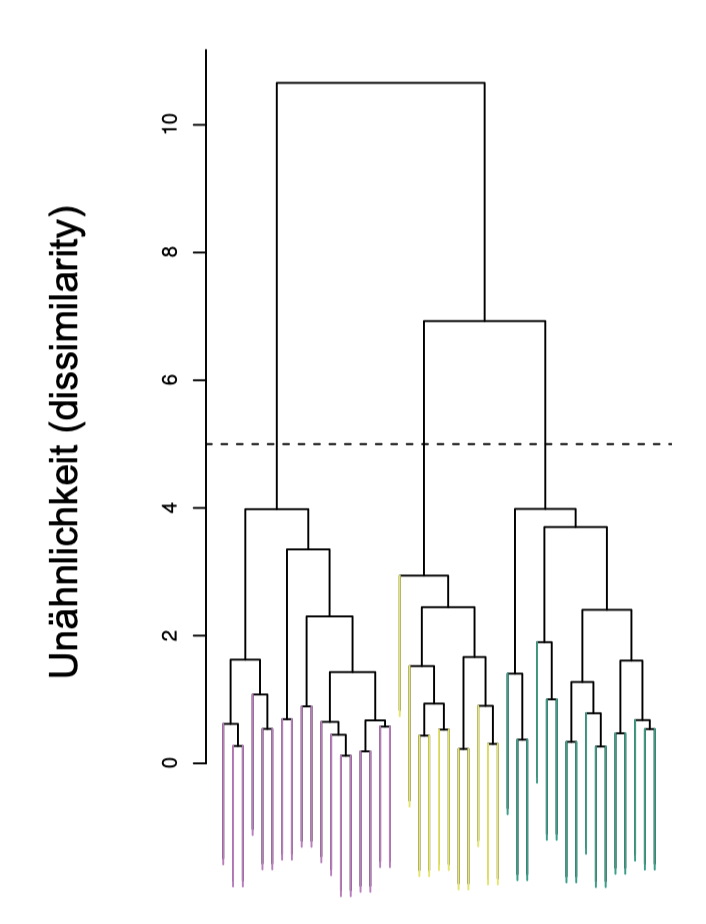
e). Drawbacks

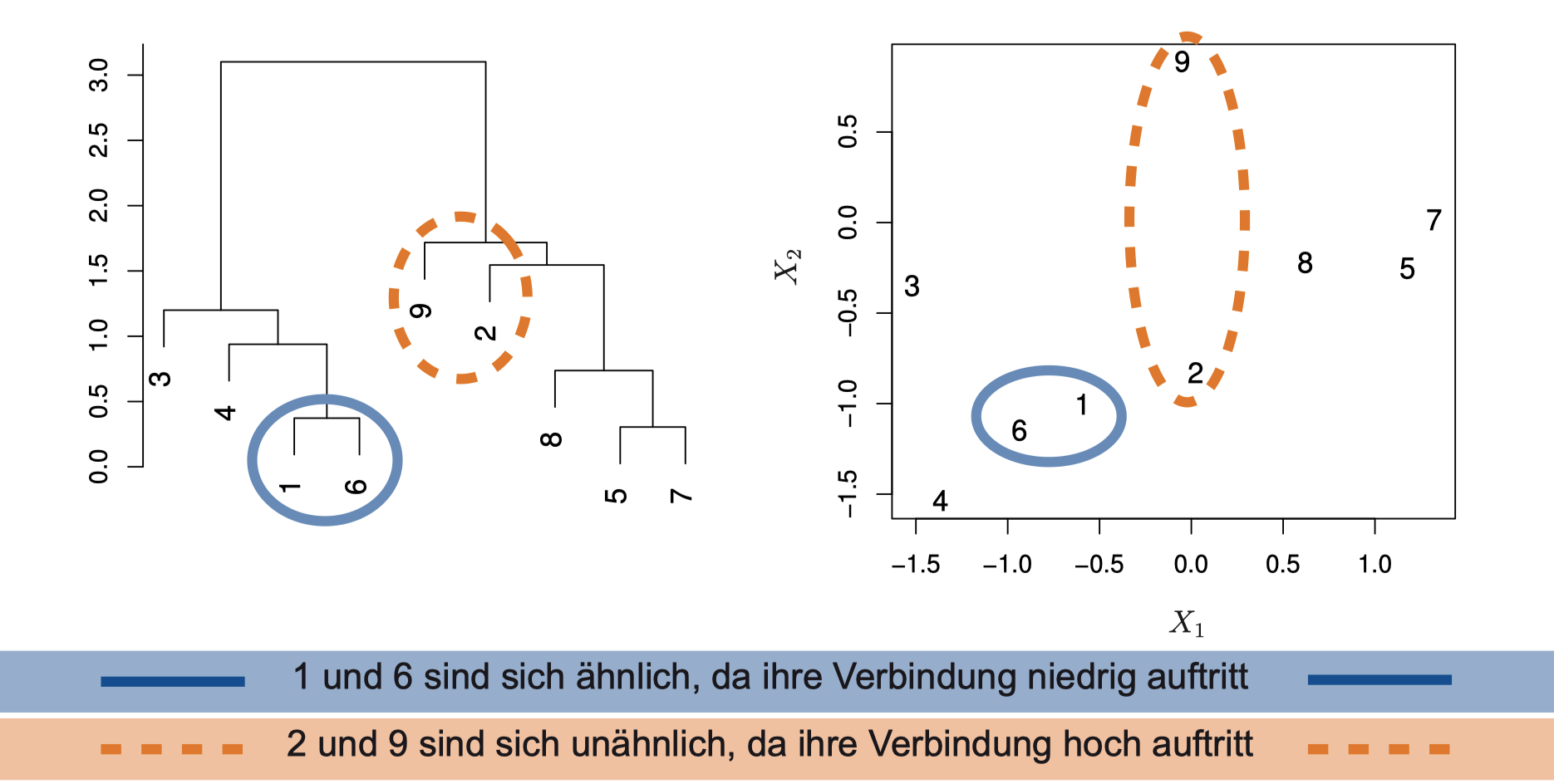


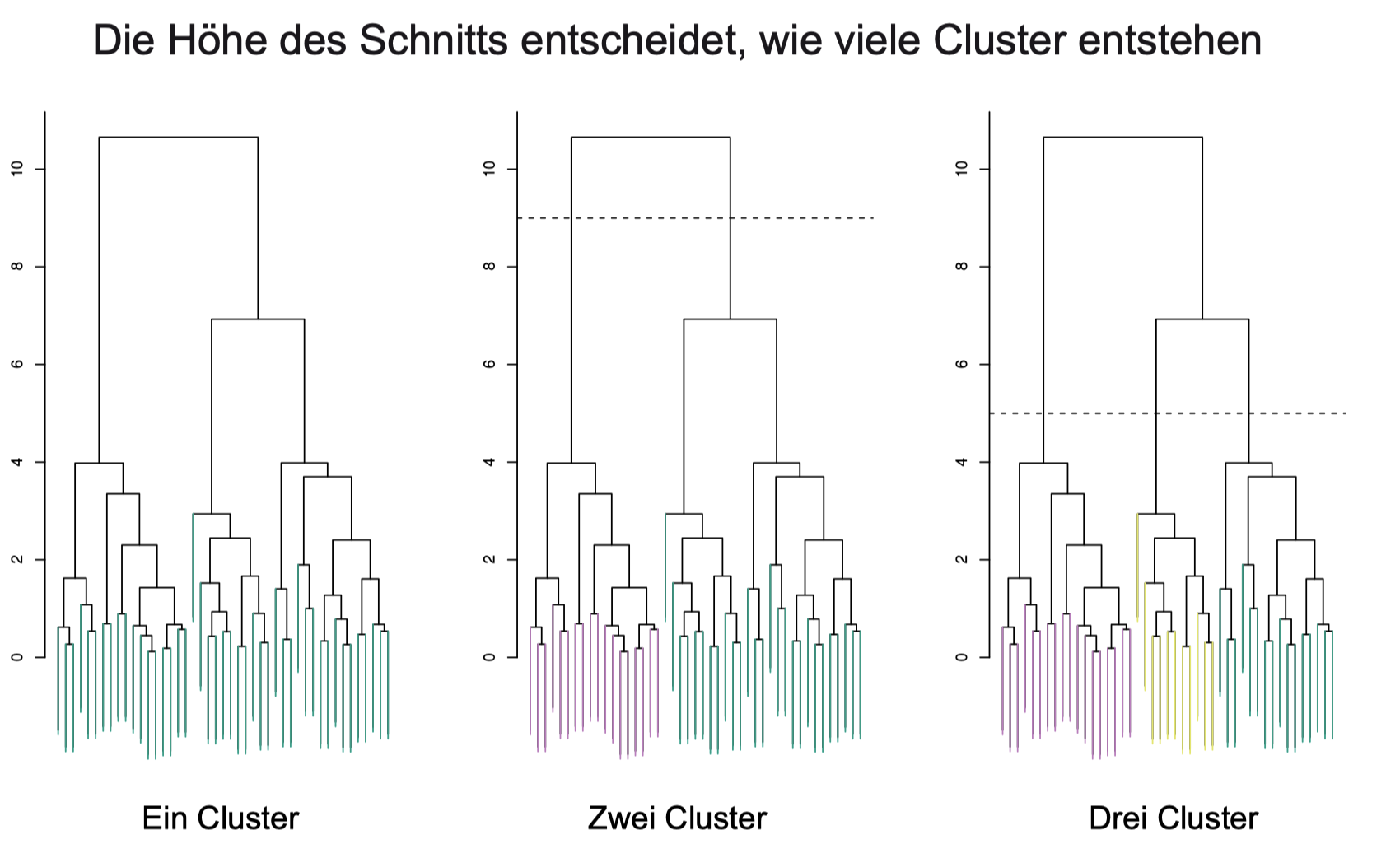
2. Hierarchical clustering



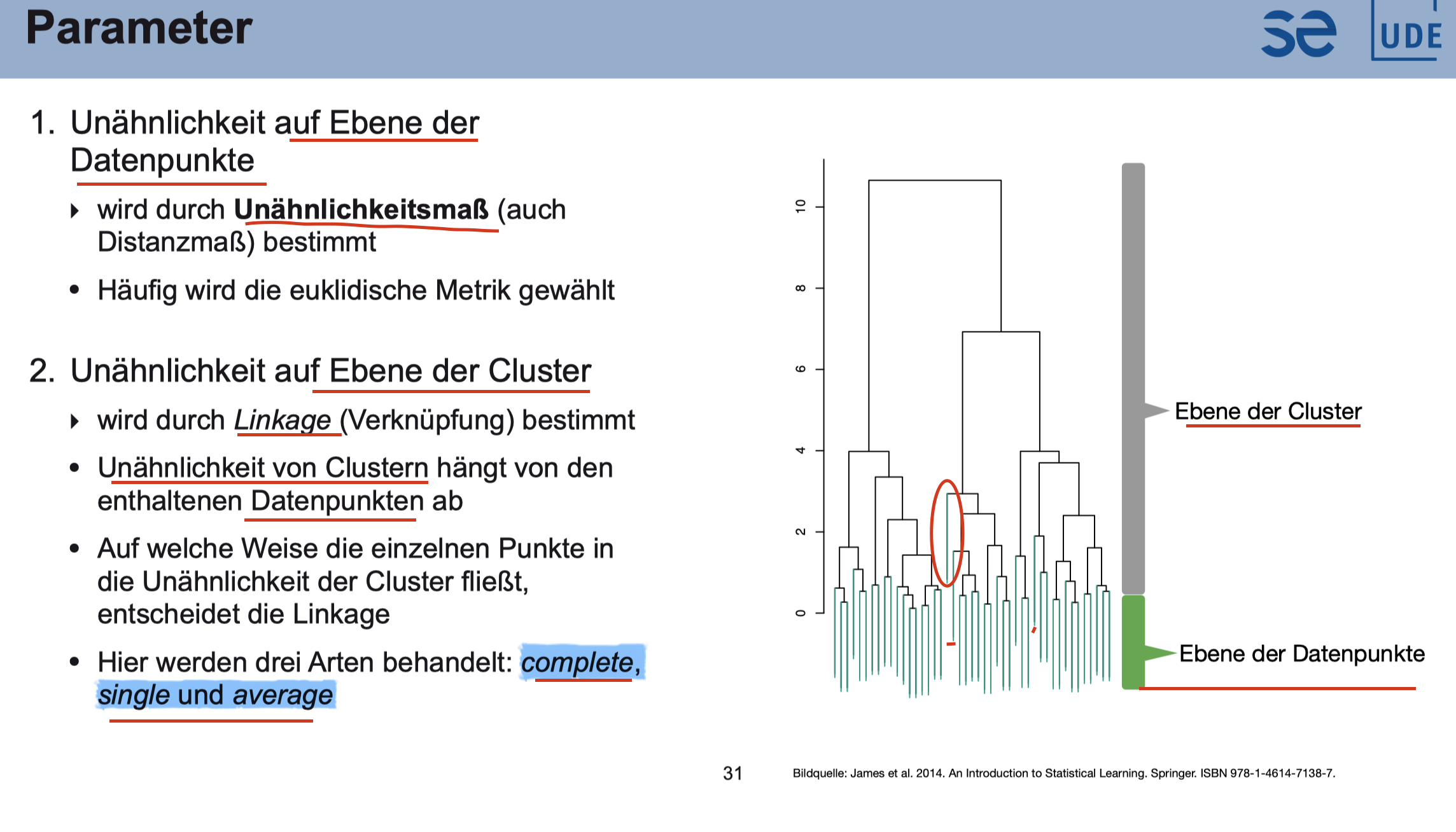
a). Dendogramm (similarity)

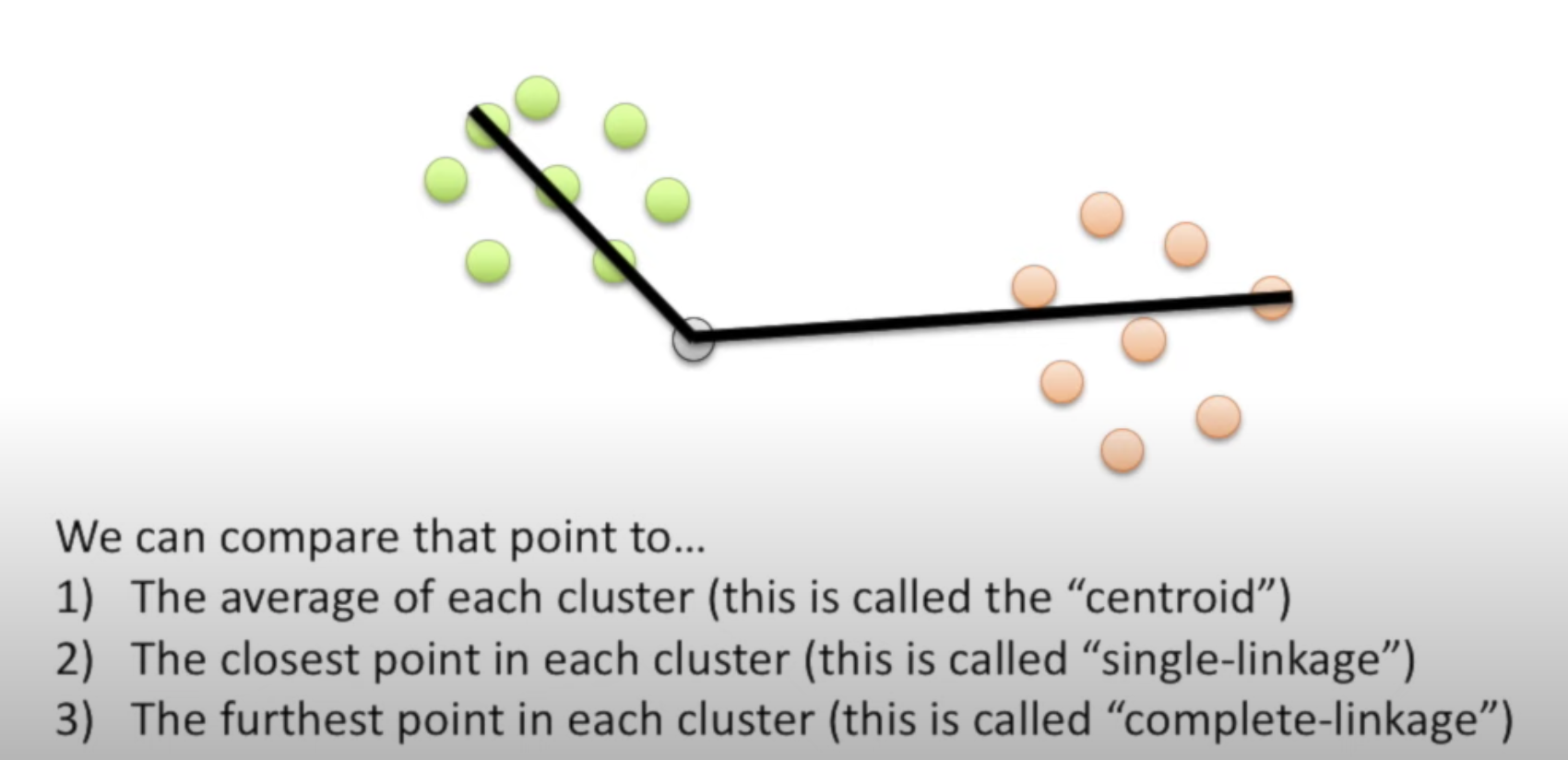


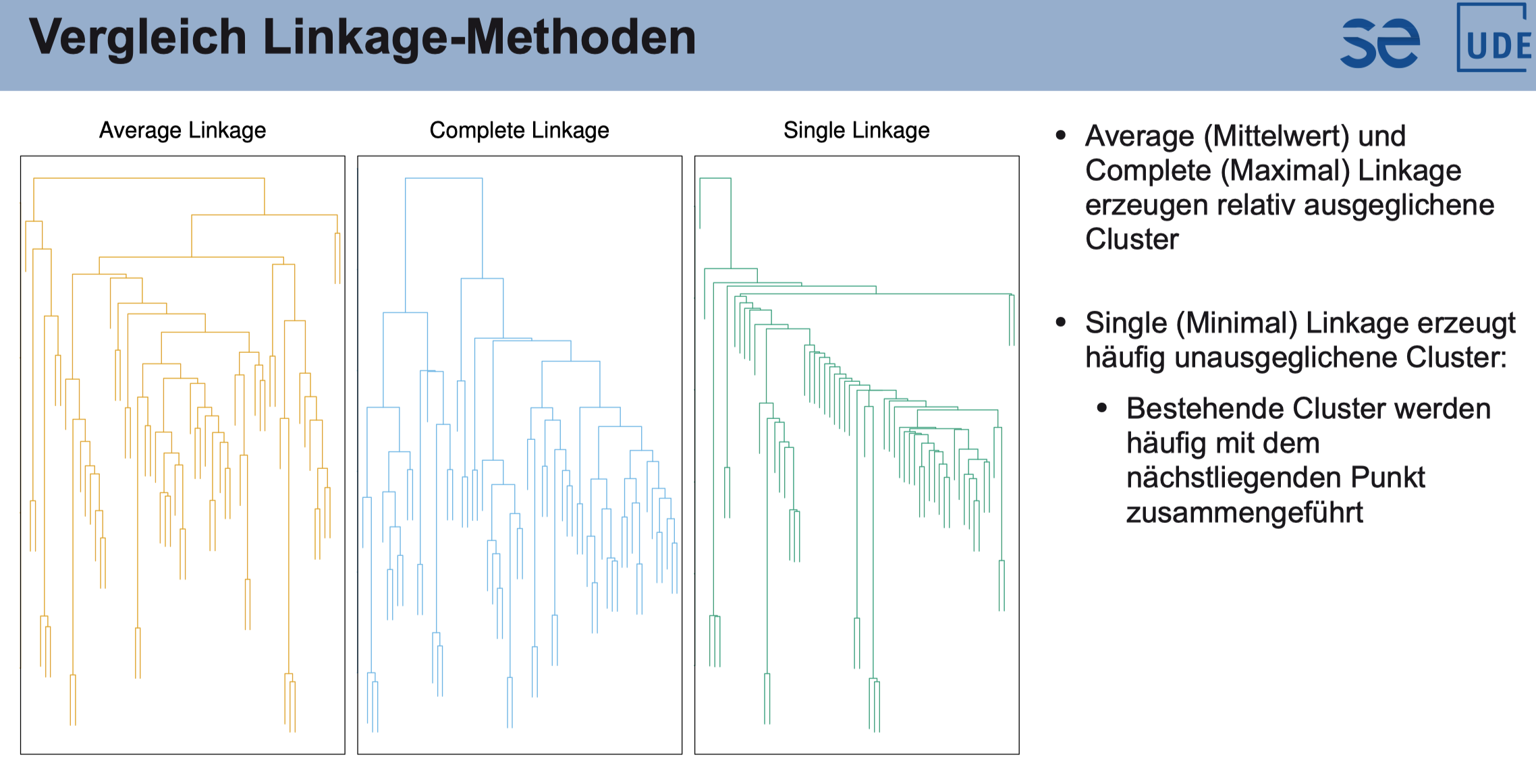




b). Linkage







c). Drawbacks:

(1). Zwingt nicht hierarchisch organisierten Daten trotzdem eine hierarchische Struktur auf

(2). Sehr rechenintensiv für hohe Dimensionen bzw. große Datenmengen (Laufzeit ist 𝒪(n3 ))

(3). Ungünstiges Zusammenführen zweier Cluster kann später nicht mehr korrigiert werden