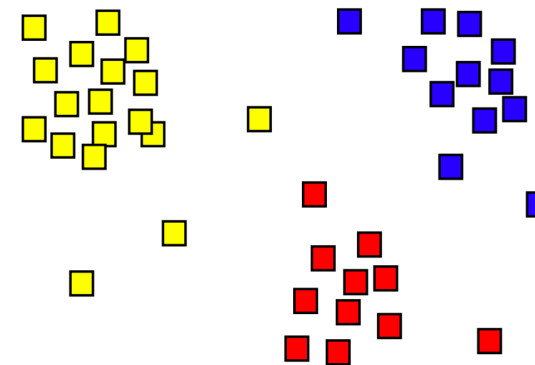
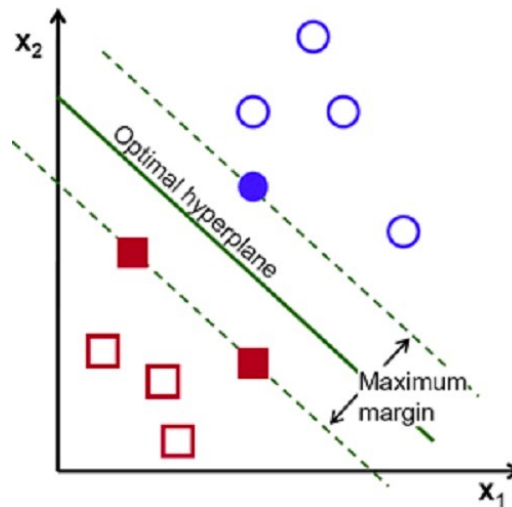


Data Mining und Maschinelles Lernen



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Wiederholung Stützvektormethode (SVM), Clustering



Basierend auf Folien von Katharina Morik, Uwe Ligges, Claus Weihs, Lutz Plümer und vielen anderen.
Danke fürs Offenlegen ihrer Folien



TECHNISCHE
UNIVERSITÄT
DARMSTADT

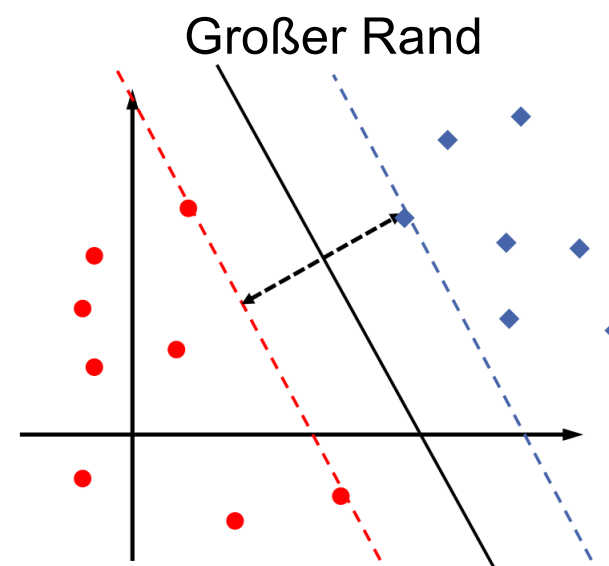
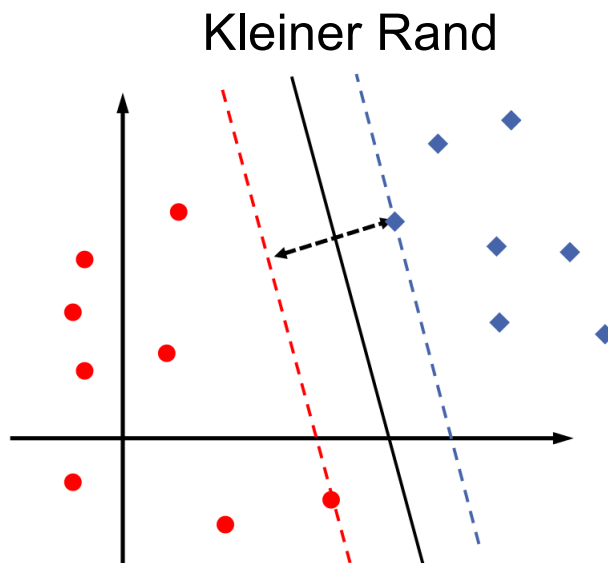


Wiederholung: Support Vector Machine (SVM)

Problem: Klassifikation, trenne die Datenmengen in zwei Partitionen

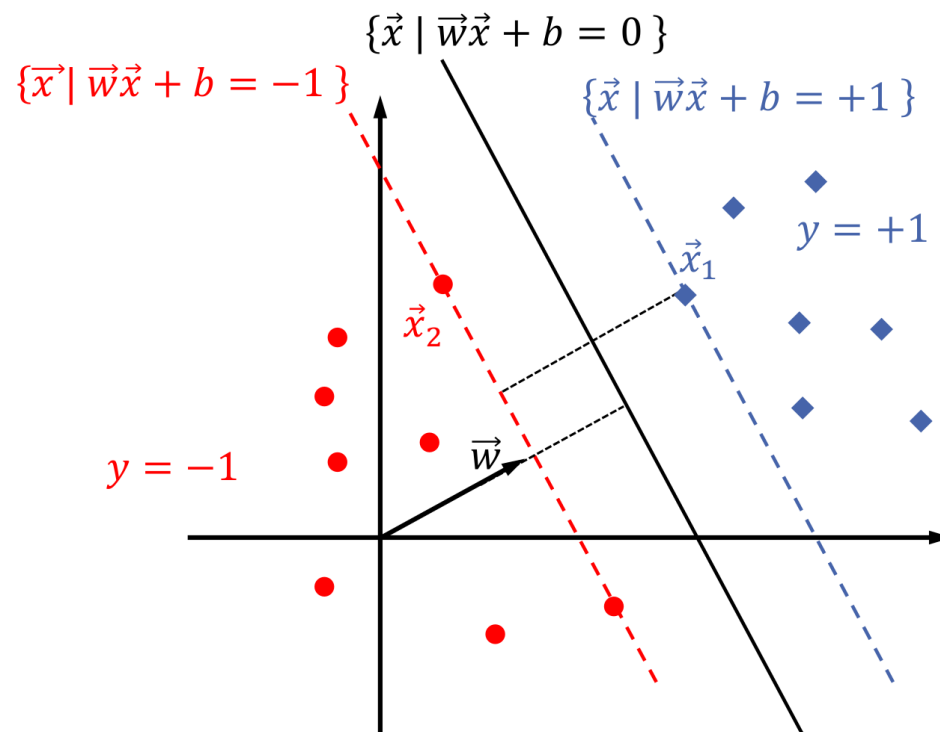
Ansatz: Finde eine optimale Trenn-Hyperebene

Intuition: Maximiere Größe des Randes (Margin) Generalisierung



Wiederholung: Support Vector Machine (SVM)

Ziel: Finde Hyperebene $\{\vec{x} \mid \vec{w}\vec{x} + b = 0\}$ mit maximalem Rand



Rand: Abstand zwischen nächsten Punkten

$$\frac{\vec{w}(\vec{x}_1 - \vec{x}_2)}{|\vec{w}|}$$

Normierung: $\vec{w}\vec{x}_1 + b = +1$
 $\vec{w}\vec{x}_2 + b = -1$

Distanz zwischen Hyperebenen:

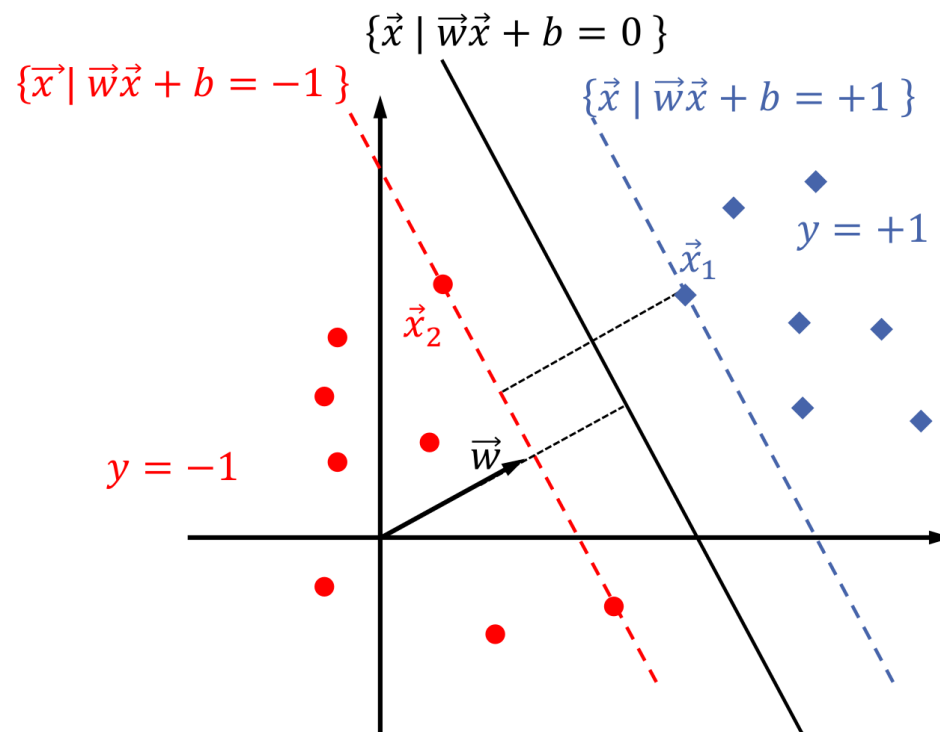
$$\vec{w}(\vec{x}_1 - \vec{x}_2) = 2$$
$$\frac{\vec{w}}{|\vec{w}|}(\vec{x}_1 - \vec{x}_2) = \frac{2}{|\vec{w}|}$$

-> Maximiere Rand

minimiere $|\vec{w}|$

Wiederholung: Support Vector Machine (SVM)

Ziel: Finde Hyperebene $\{\vec{x} \mid \vec{w}\vec{x} + b = 0\}$ mit maximalem Rand



Zielfunktion:

Maximiere $\frac{2}{|\vec{w}|}$ Minimiere $|\vec{w}|^2$

Nebenbedingung:

Alle Trainingsdaten werden korrekt klassifiziert

$$y_i(\vec{w} \vec{x}_i + b) \geq 1, \quad i = 1..n$$

Lagrange-Optimierung:

$$L_P = L(\vec{w}, b, \vec{\alpha}) = \frac{1}{2} |\vec{w}|^2 - \sum_{i=1}^n \alpha_i (y_i(\vec{w} \vec{x}_i + b) - 1)$$

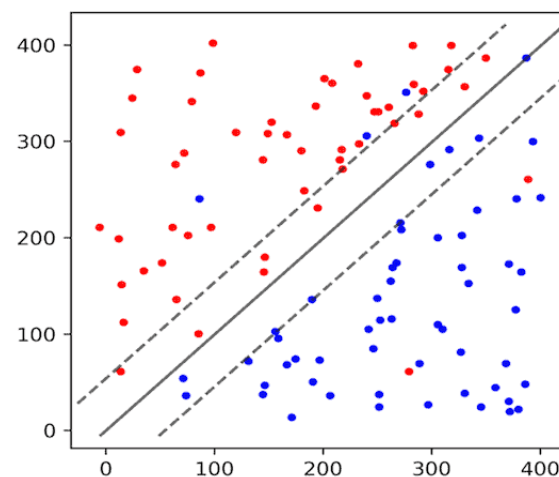
$$\vec{\alpha} = (\alpha_1, \dots, \alpha_n), \quad \alpha_1, \dots, \alpha_n \geq 0$$

Wiederholung: Support Vector Machine (SVM)

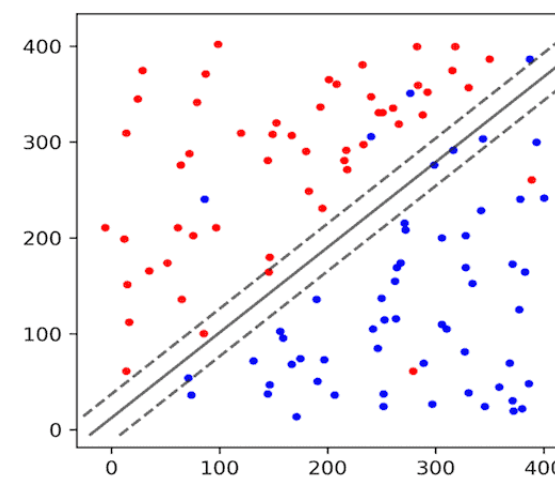
Probleme von Hard-Margin Ansatz:

- Funktioniert nur bei vollständig linear separierbaren Daten
- Anfällig gegenüber Ausreißern

Lösungsansatz: Erlaube Fehlklassifikation bei der Optimierung, aber "bestrafe" diese mit Kosten C



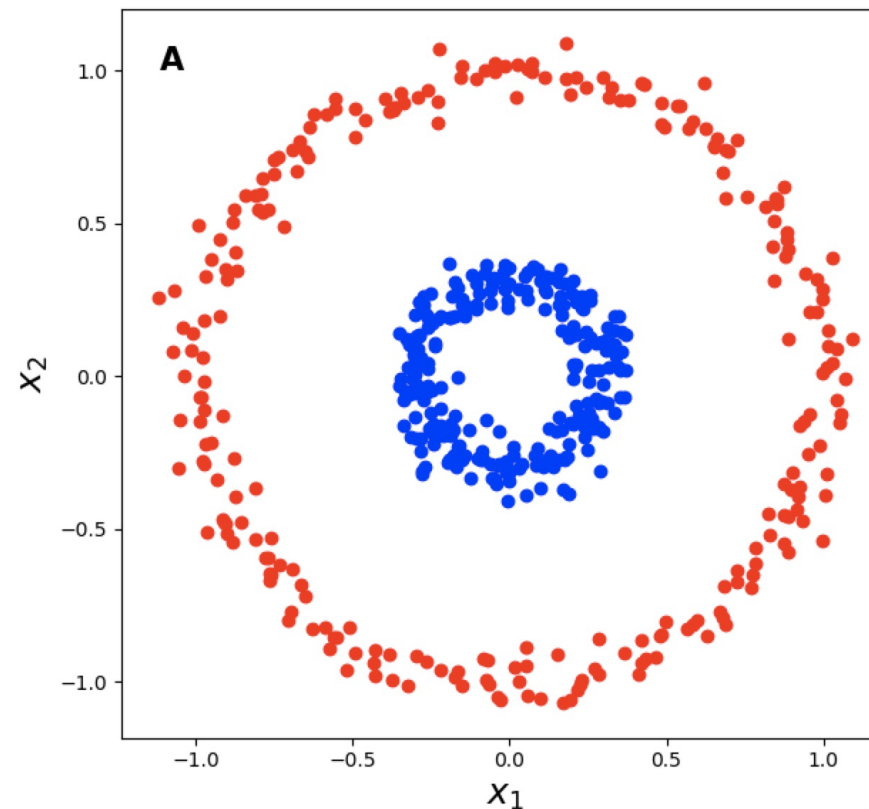
$C = 1$



$C = 100$

Wiederholung: Support Vector Machine (SVM)

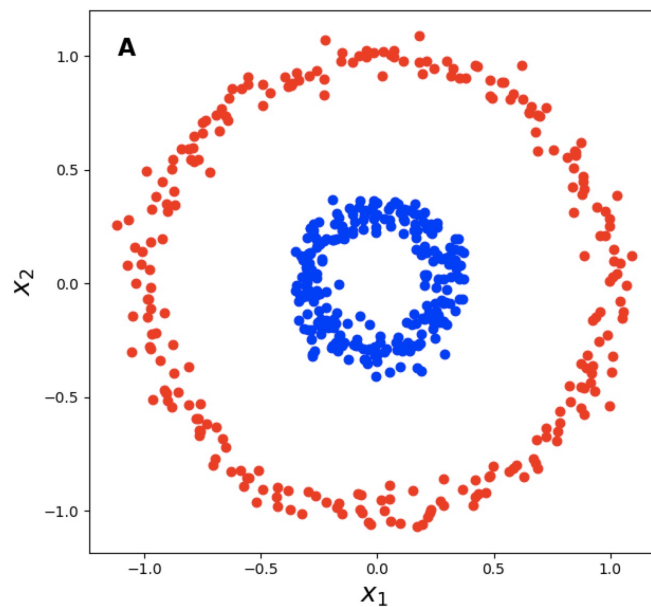
Allgemeines Problem: Nicht linear-separierbare Daten



Wiederholung: Support Vector Machine (SVM)

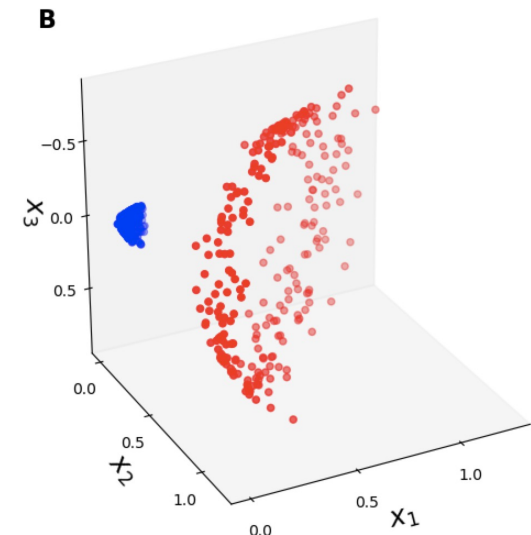
Allgemeines Problem: Nicht linear-separierbare Daten

Idee: Transformiere Daten in anderen Raum und separiere dort linear



$$\varphi(\mathbf{x}) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{bmatrix}$$

Polynomial Kernel



Wiederholung: Support Vector Machine (SVM)

Neues Problem: Transformation und Berechnung in hoch-dimensionalen Räumen ist sehr rechenintensiv


Kernel-Trick: Erlaubt implizite Berechnung in hochdimensionalen Merkmalsräumen ohne die Daten jemals explizit zu transformieren.

Kernel-Funktion: $K(\vec{x}, \vec{y}) = \phi(\vec{x})\phi(\vec{y})$

Funktion, die zwei Inputs aus dem Original-Raum entgegen nimmt und das Skalarprodukt der Vektoren im hoch-dimensionalen Raum zurück liefert.

$$L(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j (x_i * x_j)$$

Duales Optimierungsproblem

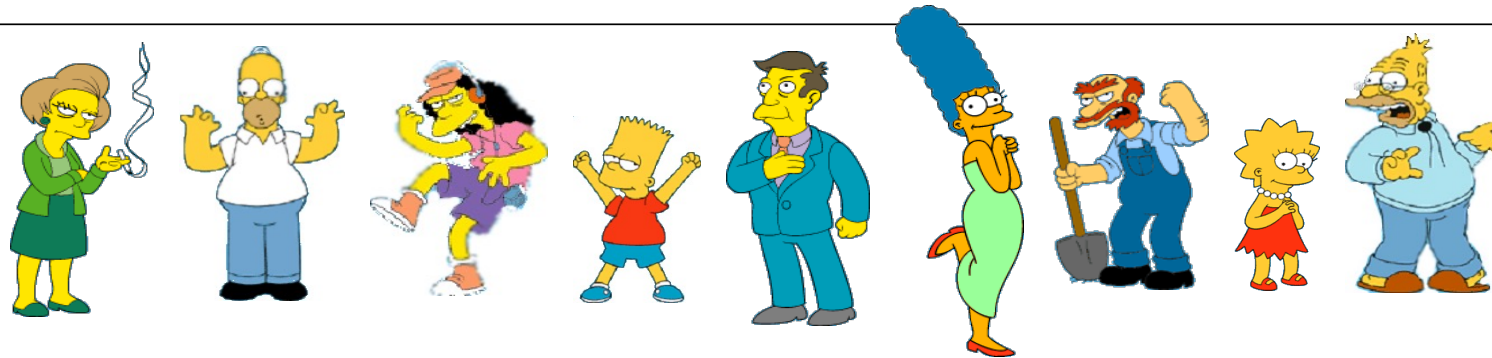

$$K(x_i, x_j) = \phi(x_i) * \phi(x_j)$$

Kernel-Funktion

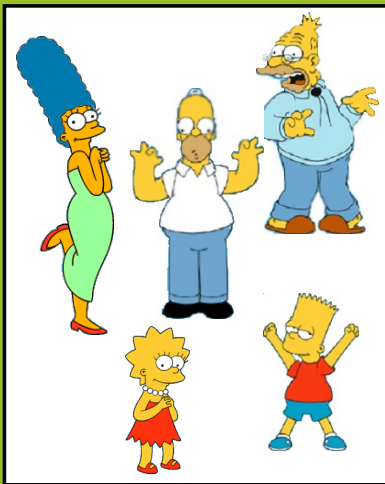
Cluster-Analyse: Wie würden Sie die Simpsons gruppieren?



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Clusteranalyse ist subjektiv



Die Simpsons



Schulangestellte



Weiblich



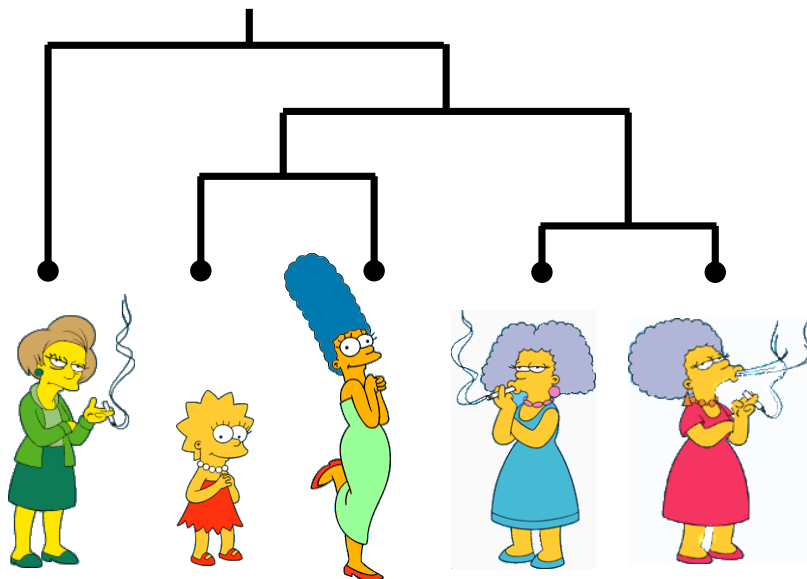
Männlich

Zwei Arten der Clusteranalyse

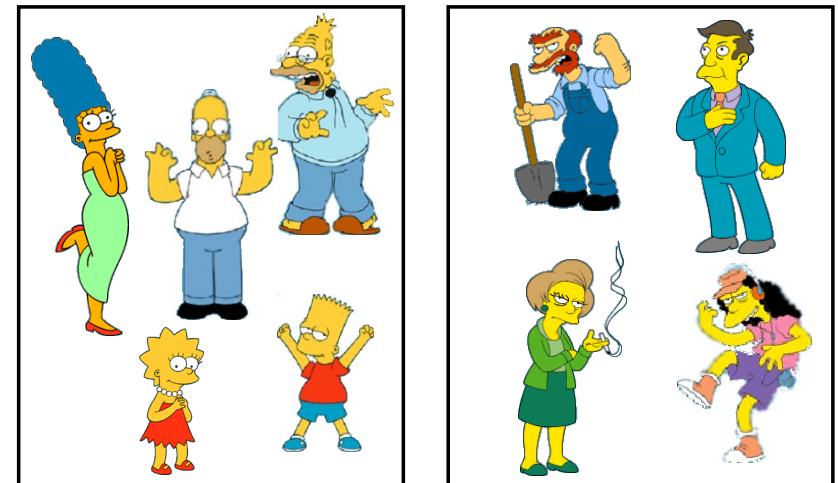
Partitionierungsansätze: Man konstruiert Partitionierungen (Aufteilungen) der Daten und bewertet sie mittels einer Bewertungsfunktion

Hierarchische Ansätze: Konstruiere eine hierarchische Aufteilung der Daten anhand eines Kriteriums

Hierarchisch



Partition



Distanzmaße

Beide Arten benötigen im Wesentlichen eine Distanzfunktion:

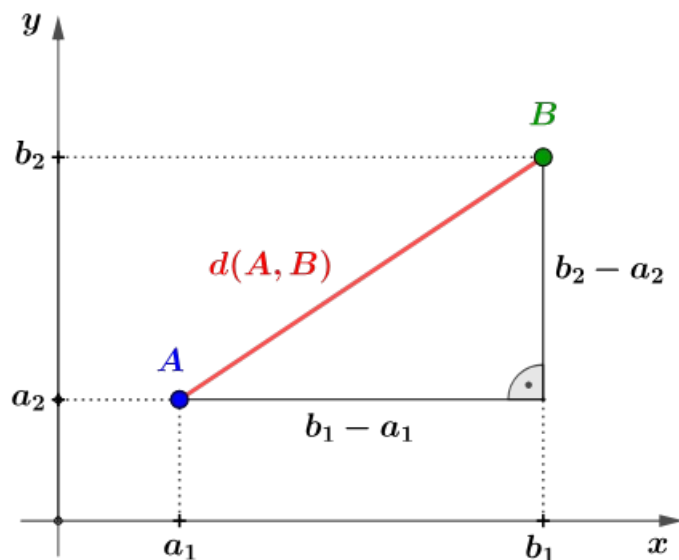
- $D(A,B) = D(B,A)$
- $D(A,A) = 0$
- $D(A,B) = 0$ If $A = B$
- $D(A,B) \leq D(A,C) + D(B,C)$

Symmetrie

Konstanz der Selbstähnlichkeit

Positive Definitheit

Dreiecksungleichung



Beispiel: Euklidische Distanz

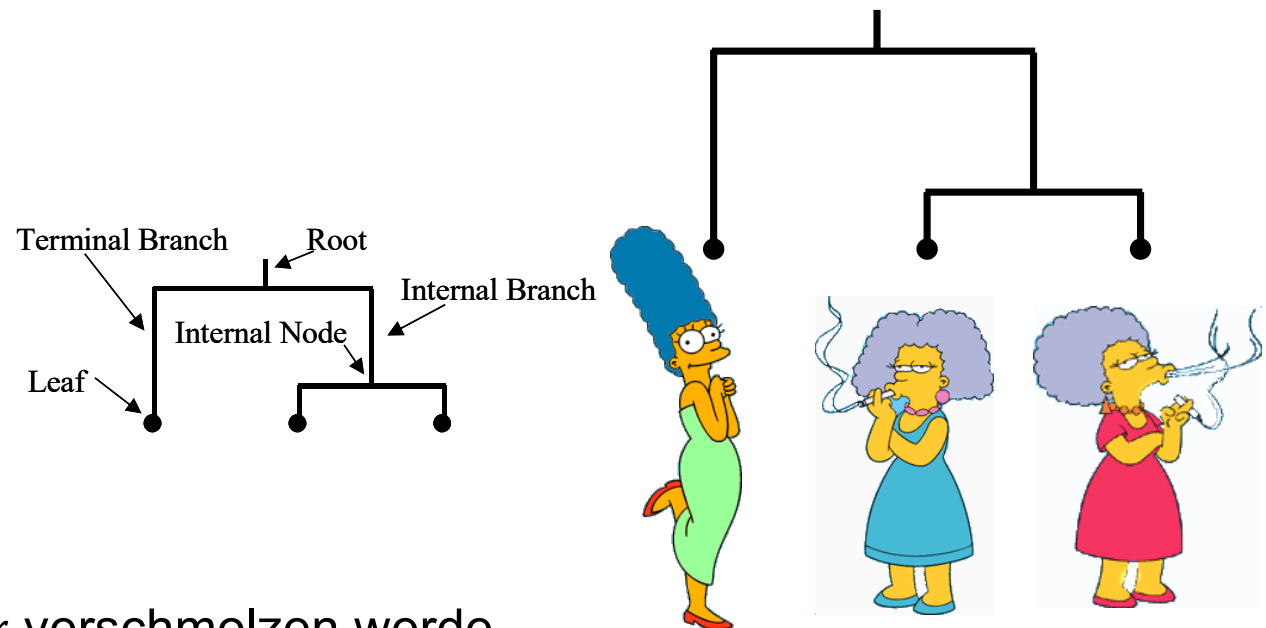
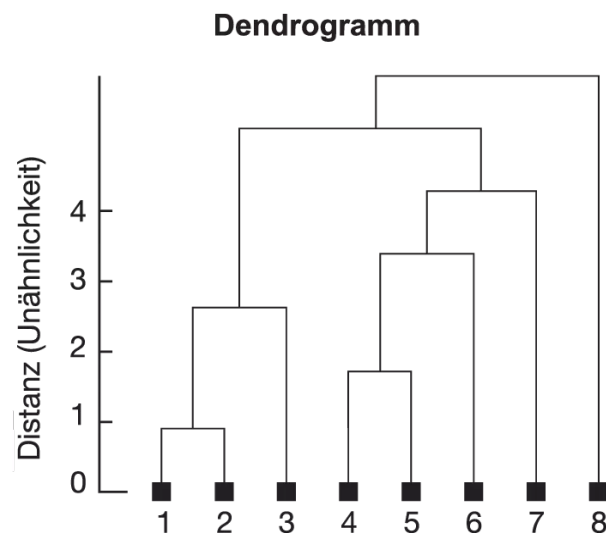


"We know it when we see it"

Hierarchisches Clustering: Dendrogramme

Baumdiagramm zur Darstellung von hierarchischem Clustering

Die Ähnlichkeit zweier Objekte wird in einem Dendrogramm durch die Höhe (von den Blättern aus gesehen) des niedrigsten internen Knoten ausgedrückt, den beide Objekte gemeinsam haben



Wenn zwei Cluster auf Höhe x verschmolzen werden, dann war der Abstand zwischen den Clustern x



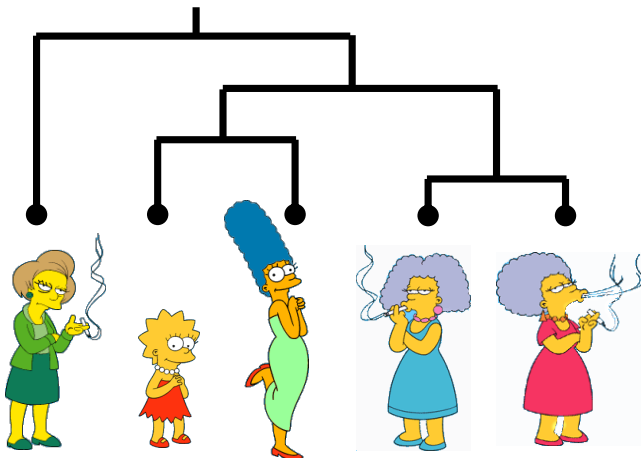
Hierarchisches Clustering: Dendrogramme

Murtagh: Counting dendrograms: A survey. Discrete Applied Mathematics 7(2):191-199 1984

Die Zahl der Dendrogramme mit n
Blättern

$$= (2n - 3)! / [(2^{n-2}) (n - 2)!]$$

Zahl der Blätter	Zahl der möglichen Dendrogramme
2	1
3	3
4	15
5	105
...	...
10	34,459,425



Zahl der Dendrogramme steigt sehr schnelle. Weil wir nicht alle durchtesten können, müssen wir uns auf Heuristiken beschränken:

Bottom-Up (Agglomerativ): Anfangs ist jedes Objekt sein eigenes Cluster. Finde die beiden Cluster, die sich am ähnlichsten sind, und vereinige (merge) sie. Wiederhole das solange, bis es nur noch ein Cluster gibt.

Top-Down (Aufteilend): Anfangs sind alle Objekte in einem Cluster. Finde den besten Split und führe diesen aus. Wiederhole das so oft, bis die Cluster nur noch aus einem Objekt bestehen

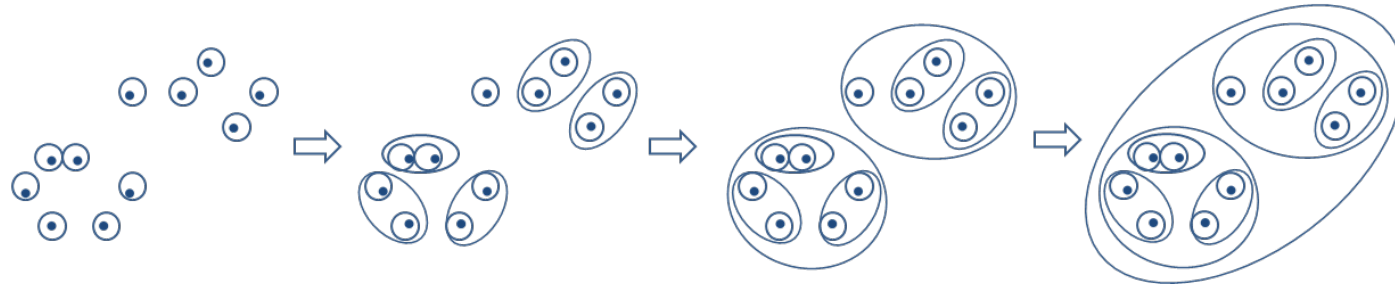


Hierarchisches Clustering: Dendrogramme

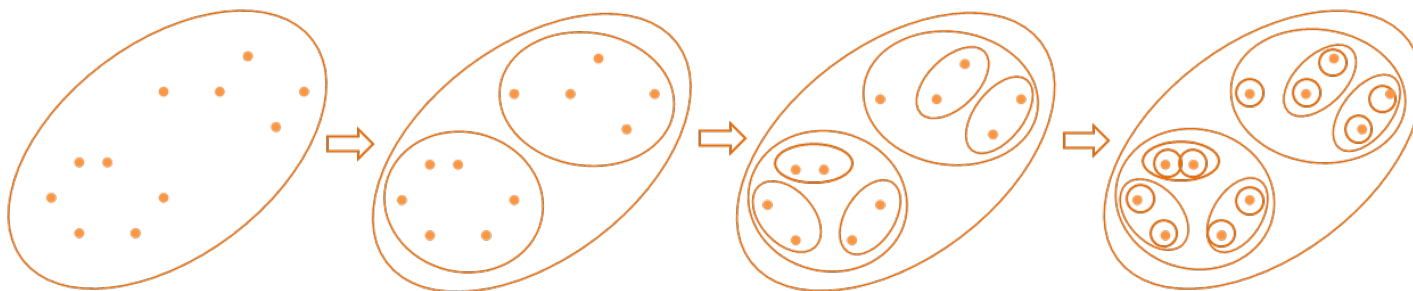
Bottom-Up (Agglomerativ): Anfangs ist jedes Objekt sein eigenes Cluster. Finde die beiden Cluster, die sich am ähnlichsten sind, und vereinige (merge) sie. Wiederhole das solange, bis es nur noch ein Cluster gibt.

Top-Down (Aufteilend): Anfangs sind alle Objekte in einem Cluster. Finde den besten Split und führe diesen aus. Wiederhole das so oft, bis die Cluster nur noch aus einem Objekt bestehen

Agglomerative Hierarchical Clustering



Divisive Hierarchical Clustering



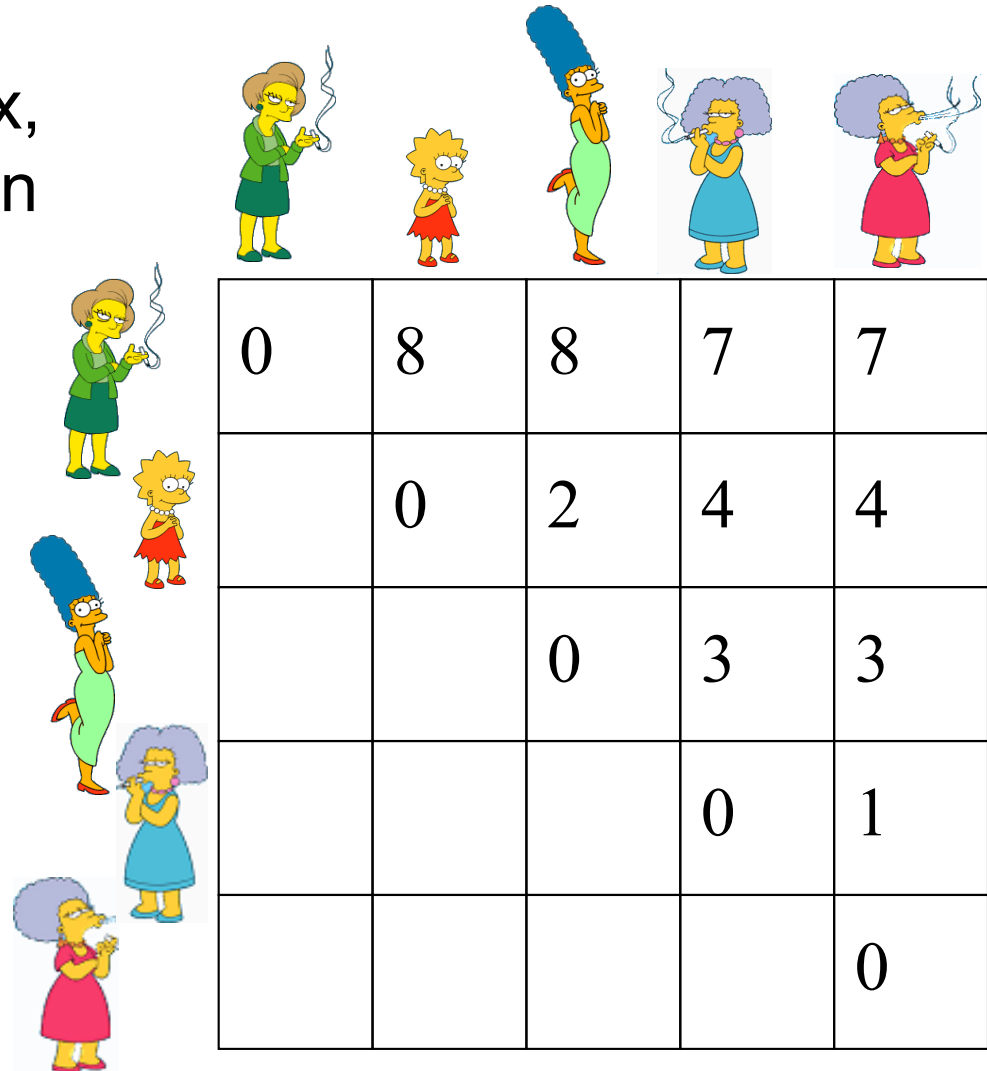
Bildquelle: <https://www.pinecone.io/learn/k-means-clustering/>

Beispiel Bottom-Up

Wir haben eine Distanzmatrix, die alle paarweisen Distanzen enthält

$$D(\text{Marge Simpson}, \text{Lisa Simpson}) = 8$$

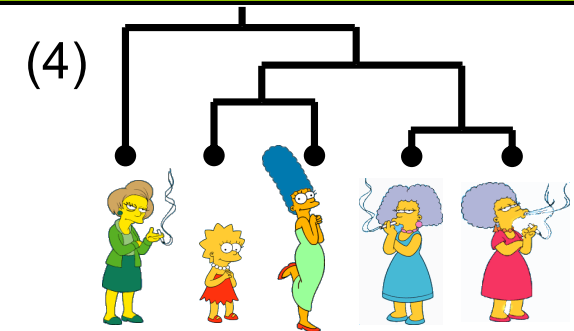
$$D(\text{Granny Simpson}, \text{Auntie Simpson}) = 1$$



0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

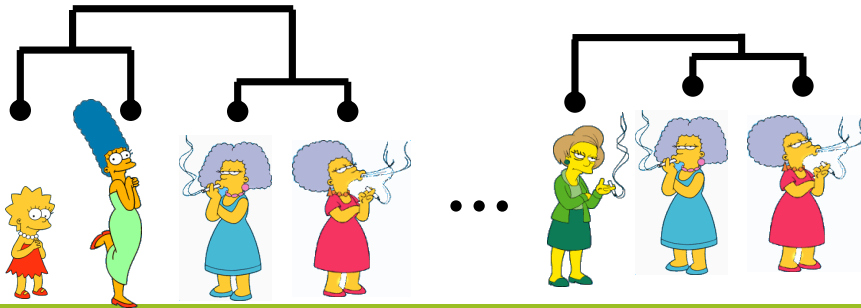
Beispiel Bottom-Up

Anfangs ist jedes Objekt sein eigenes Cluster. (1) Finde die beiden Cluster, die sich am ähnlichsten sind, und vereinige (merge) sie. Wiederhole (2,3) das solange, bis es nur noch ein Cluster gibt (4).

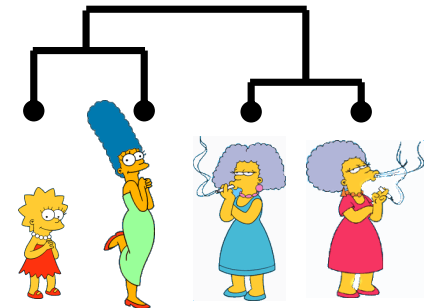


(3)

Betrachte alle
möglichen
Vereinigungen

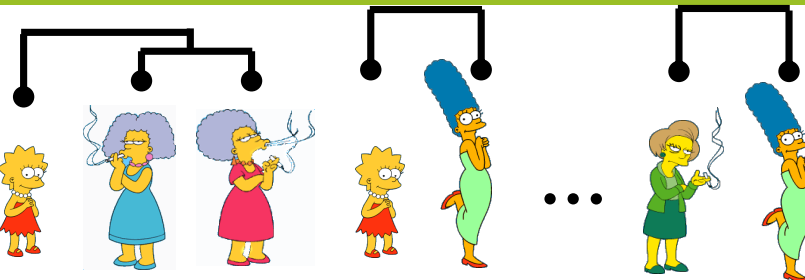


Wähle die
beste

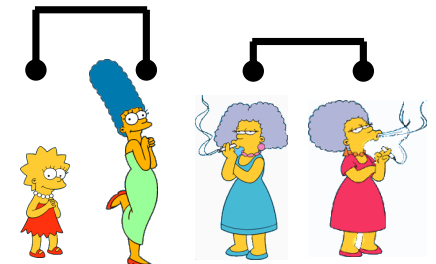


(2)

Betrachte alle
möglichen
Vereinigungen

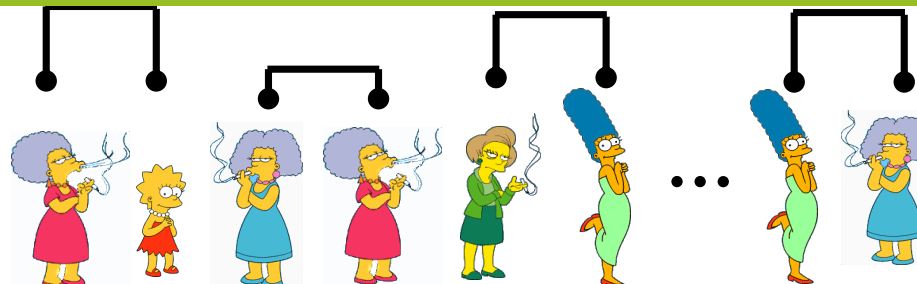


Wähle die
beste



(1)

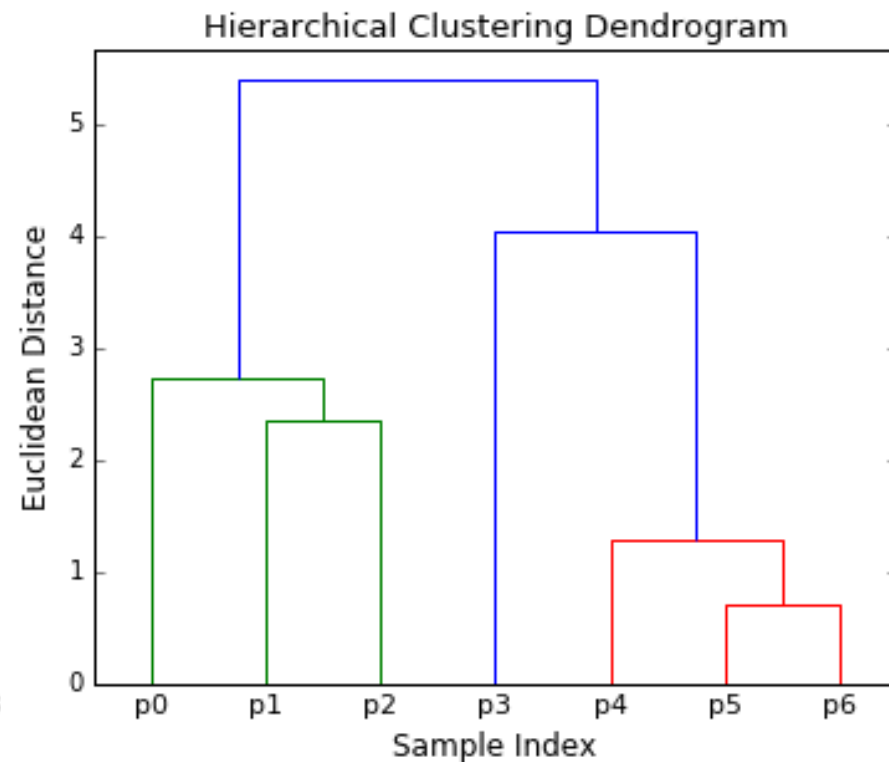
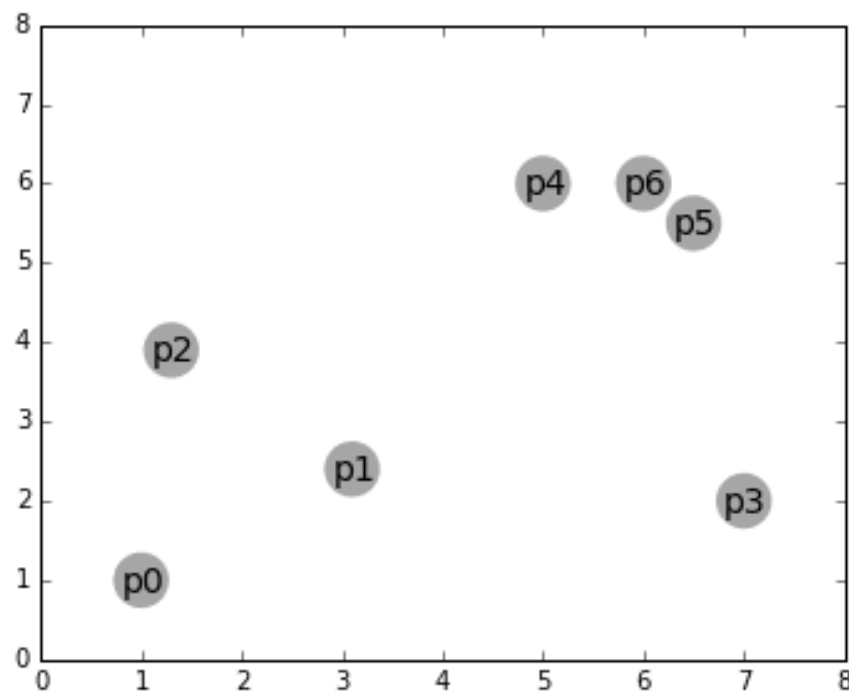
Betrachte alle
möglichen
Vereinigungen



Wähle die
beste



Hierarchisches Clustering: Dendrogramme



Bildquelle: <https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/>



Dendrograms / Hierarchisches Clustering

Pedro

Petros (Greek), Peter (English), Piotr (Polish), Peadar (Irish), Pierre (French), Peder (Danish), Peka (Hawaiian), Pietro (Italian), Piero (Italian Alternative), Petr (Czech), Pyotr (Russian)

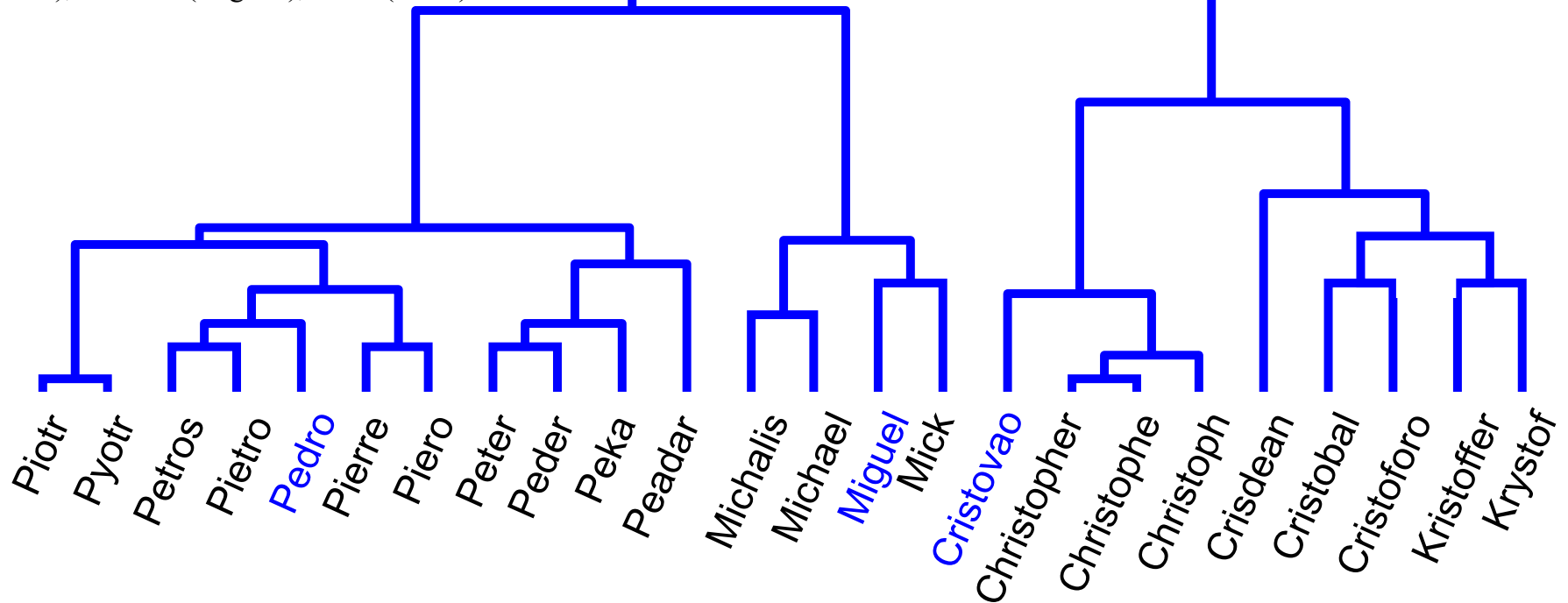
Cristovao

Christoph (German), Christophe (French), Cristobal (Spanish), Cristoforo (Italian), Kristoffer (Scandinavian), Krystof (Czech), Christopher (English)

Miguel

Michalis (Greek), Michael (English), Mick (Irish!)

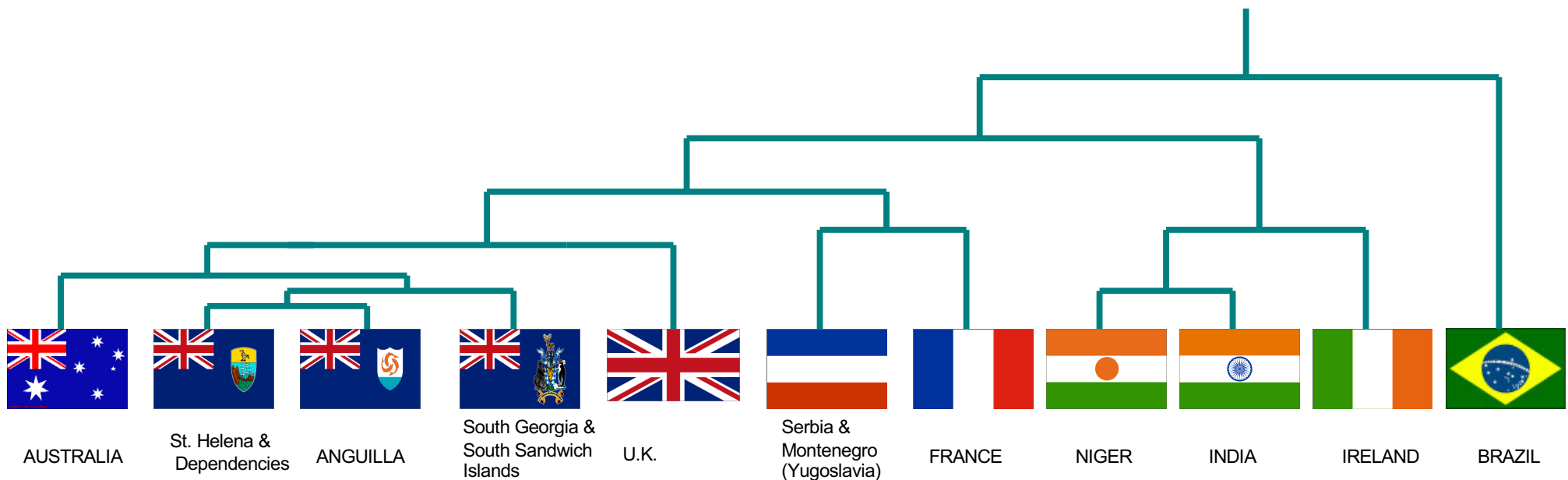
Distanz = Wieviele
Editieroperationen
brauchen wir um String
A in String B zu
überführen ?



Hierarchien können Strukturen aufdecken, aber auch vortäuschen

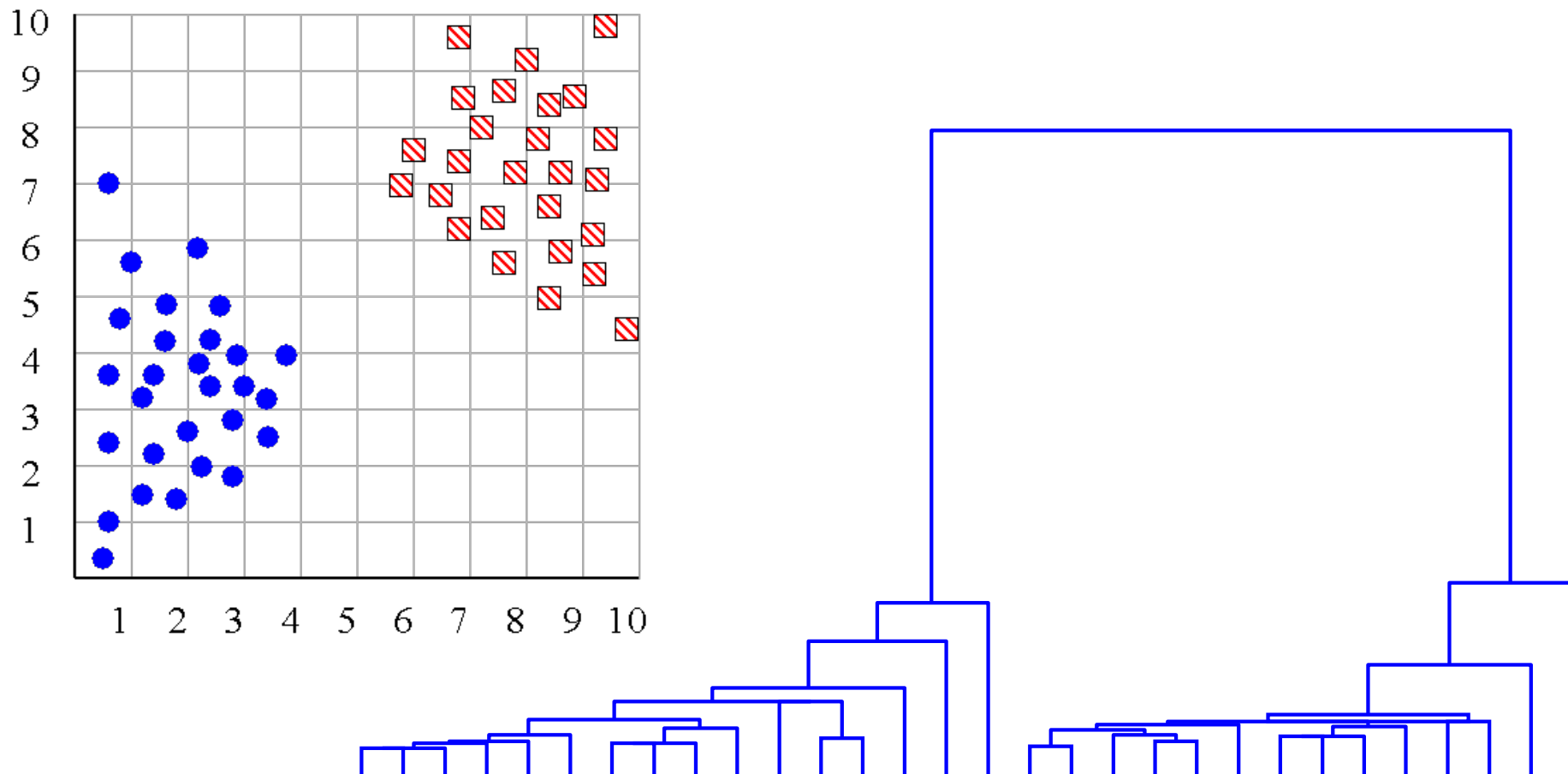
Die dichte Gruppe um Australien macht Sinn. Es sind alle Ländern aus den ehemaligen Britischen Kolonien.

Aber Nigeria und Indien (von Irland wollen wir mal gar nicht erst sprechen) haben nicht viel mit einander am Hut.



Dendrogramme können uns manchmal auch die "richtige" Anzahl an Clustern anzeigen

Die zwei stark getrennten Teilbäume legen es nahe, dass es zwei übergeordnete Cluster gibt. Normalerweise ist das aber nicht so klar!

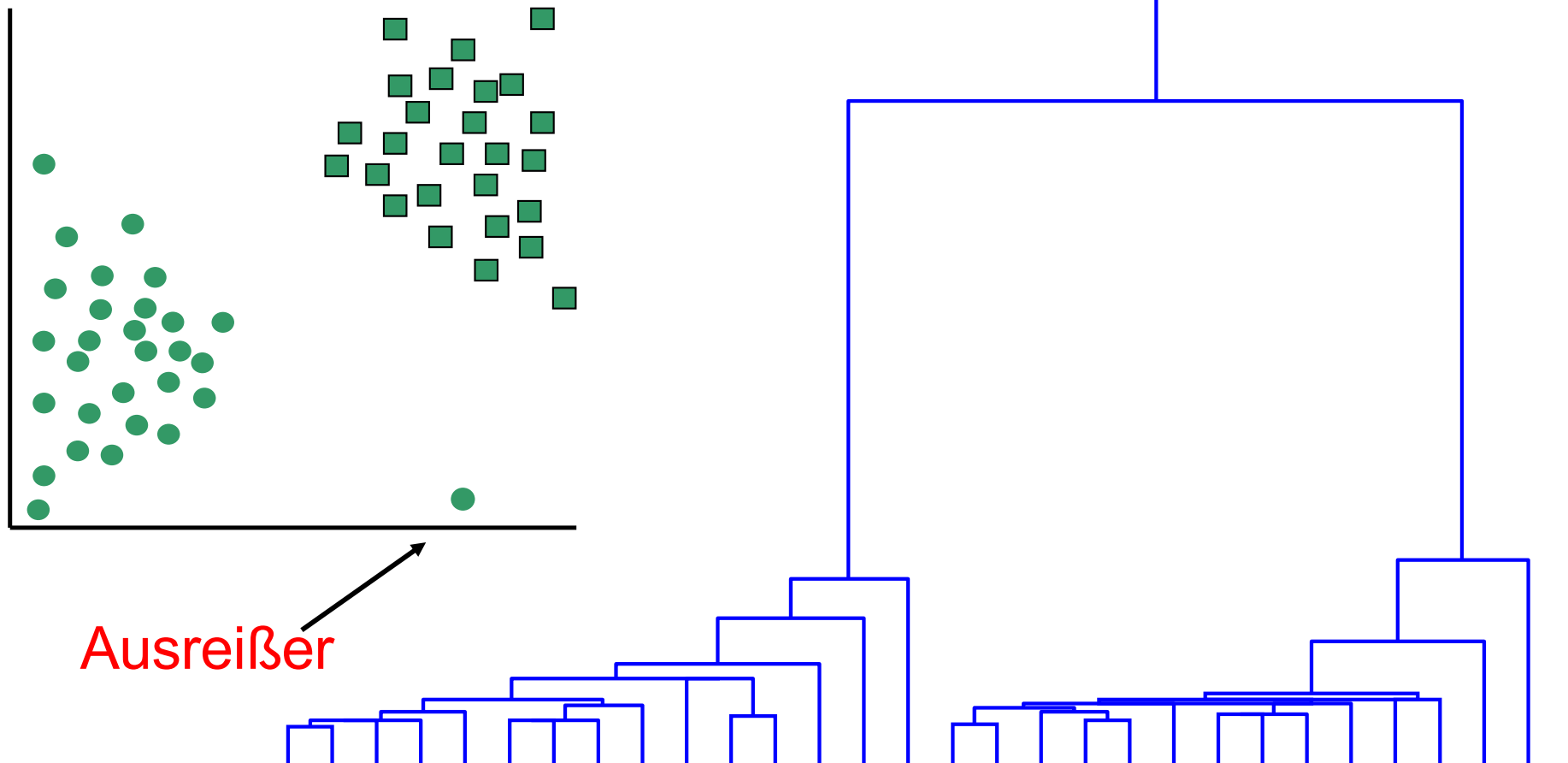


Dendrogramme können auch Ausreißer in den Daten feststellen



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Dieser einzelne Ast legt nahe, dass es sich um einen Ausreißer handelt.



TECHNISCHE
UNIVERSITÄT
DARMSTADT





Clustering ist eine Kunst

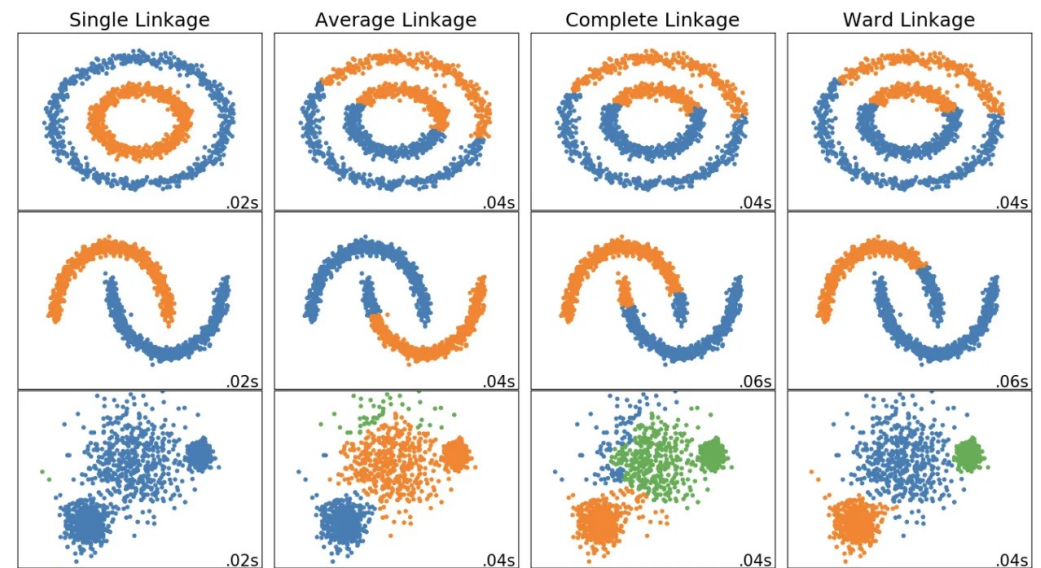
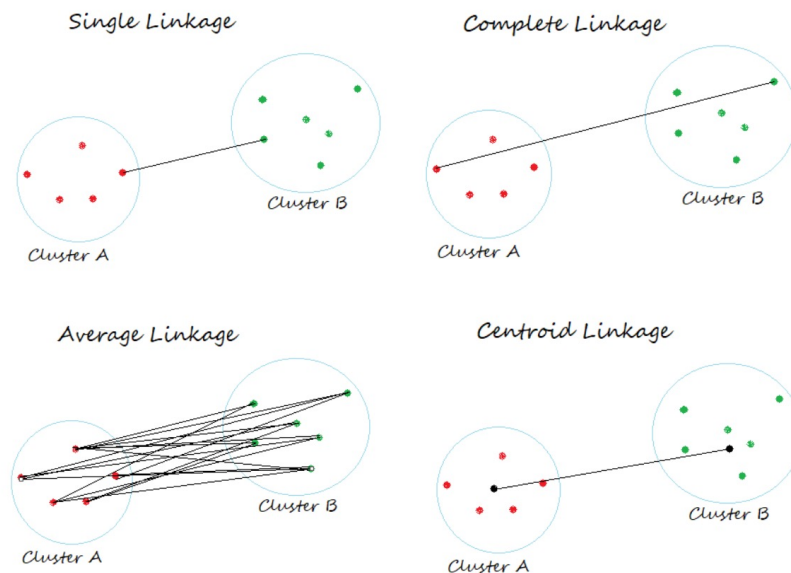
Selbst wenn wir eine gute Distanzfunktion haben, ist es nicht selbstverständlich, wie wir eine Distanz **zwischen einem Objekt und einem Cluster** bzw. **zwischen Clustern** definieren.

- **Single linkage (nearest neighbor):** Die Distanz ist die Distanz zwischen den beiden nächsten Nachbarn in den beiden unterschiedlichen Clustern.
- **Complete linkage (furthest neighbor):** Die Distanz ist die Distanz zwischen den beiden entferntesten Objekten
- **Group average linkage:** Durchschnitt aller paarweisen Distanzen
- **Wards Linkage:** Man versucht die Varianz zwischen den Clustern zu minimieren bzw. den Zuwachs an Varianz bei der Zusammenführung zweier Cluster zu minimieren.

Linkage-Kriterium kann großen Einfluss auf das Clustering Ergebnis haben



TECHNISCHE
UNIVERSITÄT
DARMSTADT



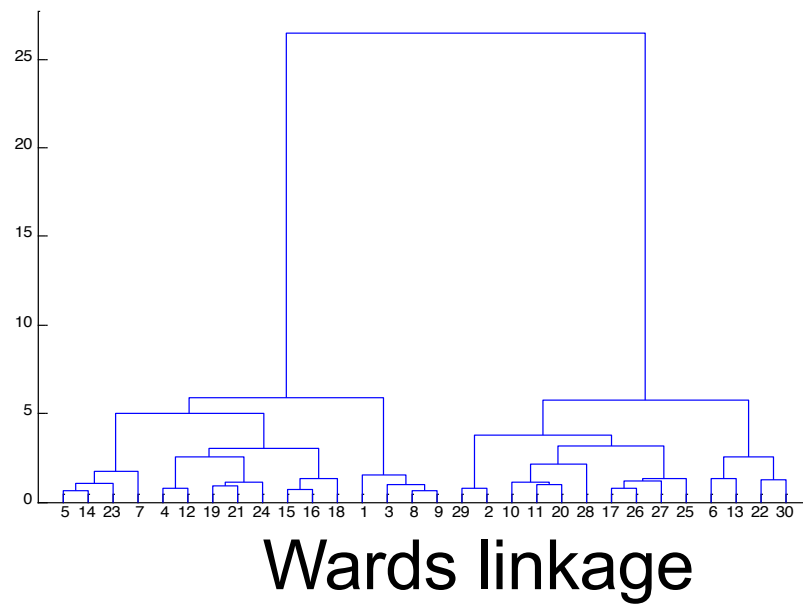
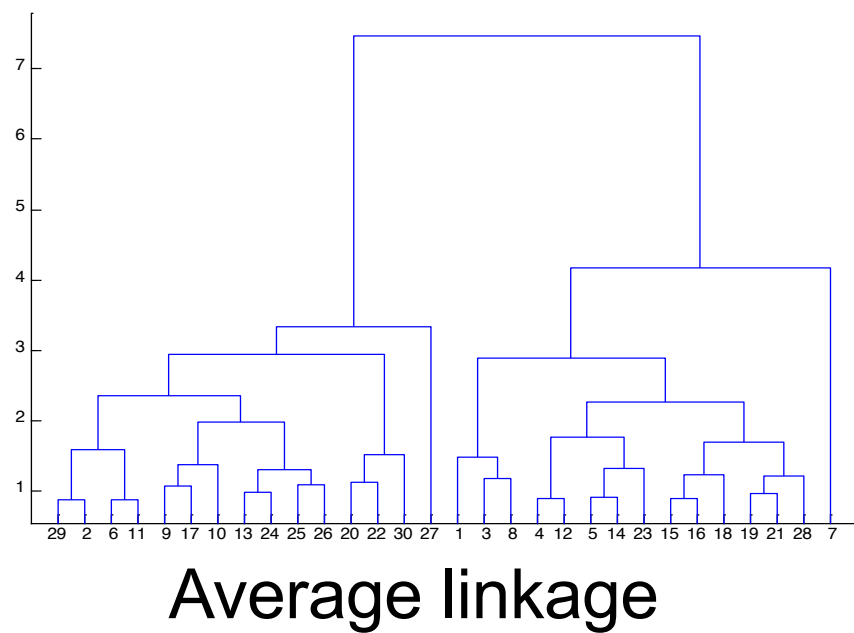
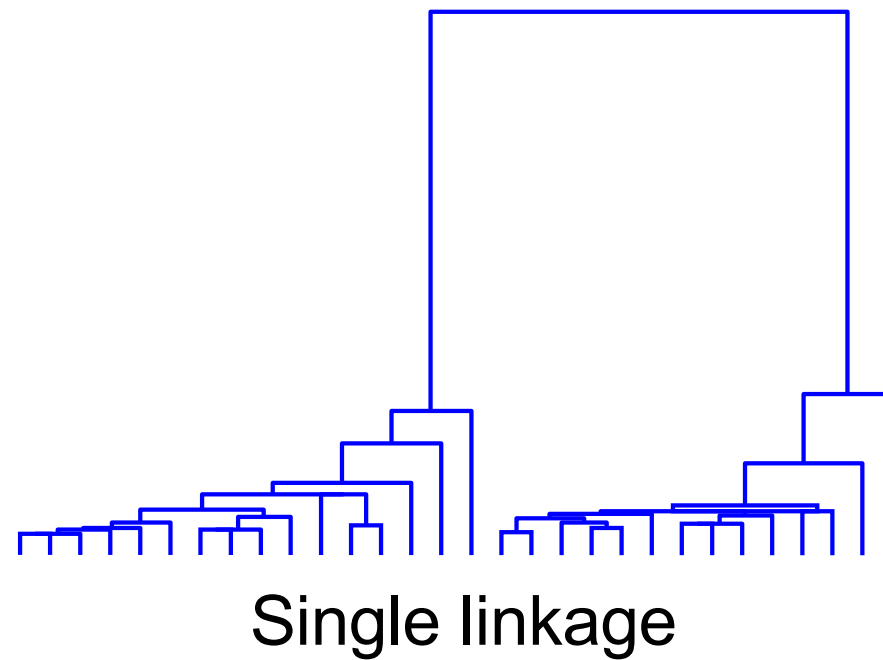
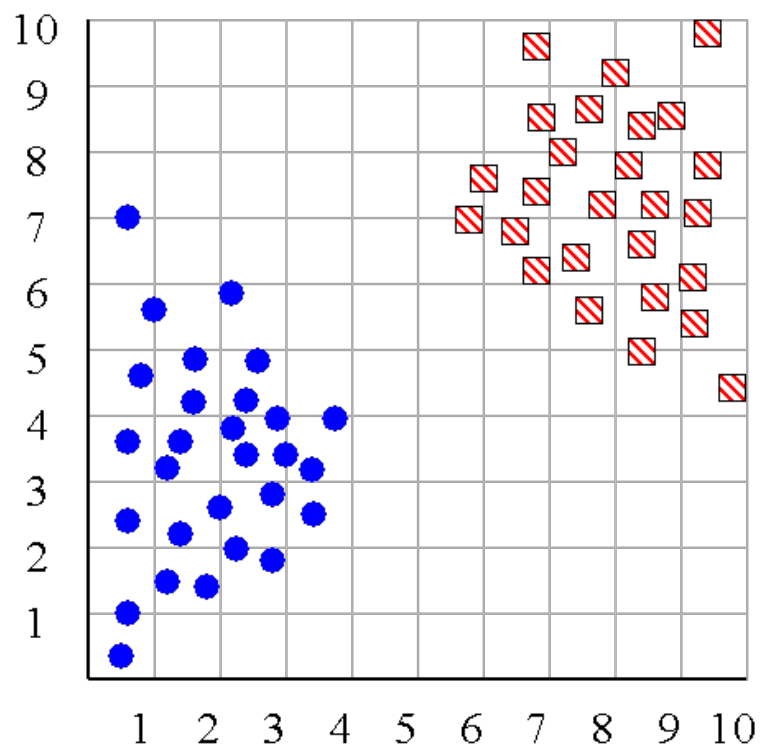
Bildquelle: <https://www.analyticsvidhya.com/blog/2021/06/single-link-hierarchical-clustering-clearly-explained/>

Bildquelle: <https://towardsdatascience.com/machine-learning-algorithms-part-12-hierarchical-agglomerative-clustering-example-in-python-1e18e0075019>



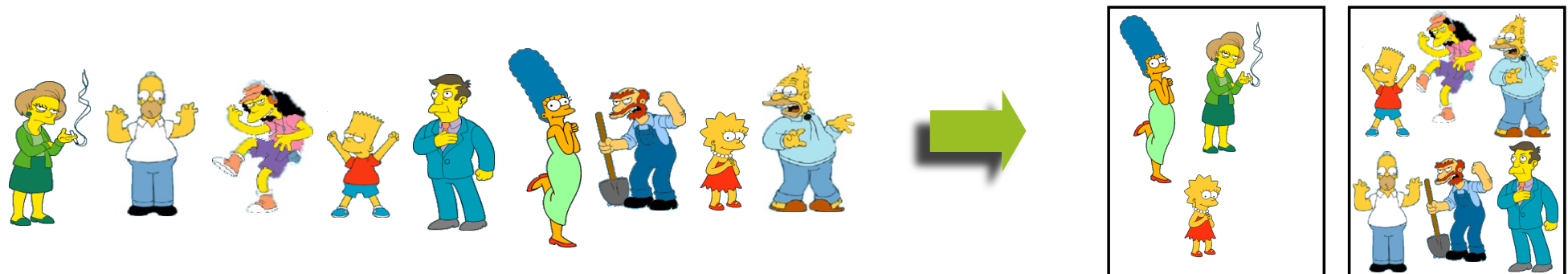
TECHNISCHE
UNIVERSITÄT
DARMSTADT





Clustering mittels Partitionierungen

Keine Hierarchie. Jedes Objekt gehört zu genau einem Cluster. Die Cluster überlappen nicht



Partitionierungsansatz: K-Means Algorithmus

Hyperparameter (muss zu Beginn festgelegt werden):

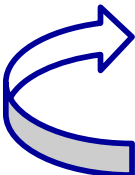
K = Anzahl an Cluster

Art der Initialisierung der Clusterzentren

Vorgehen:

(1) **Initialisiere die Clusterzentren** (z.B. indem K Datenpunkte zufällig gewählt werden)

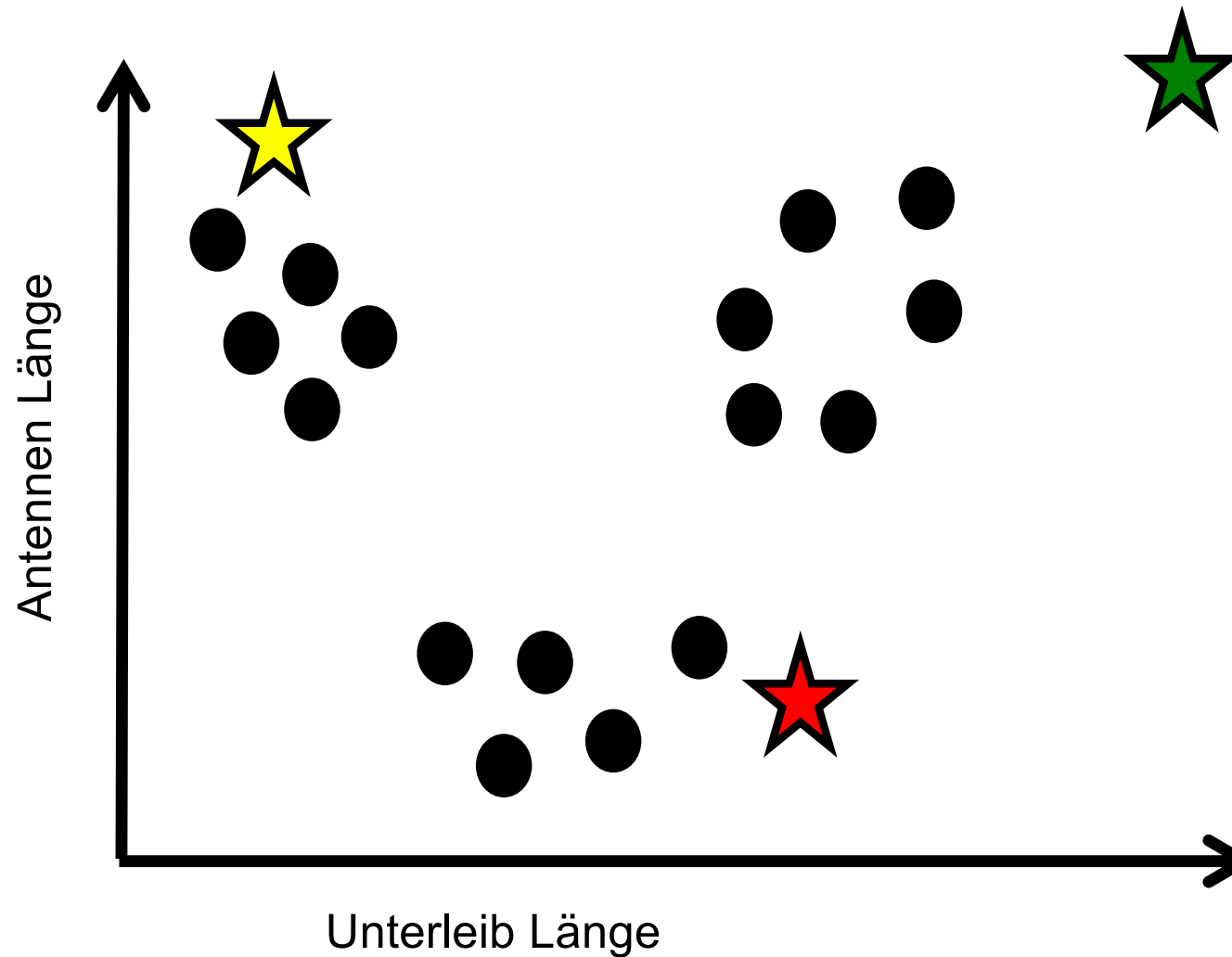
Jetzt, wiederholen wir die folgenden zwei Schritte bis zur Konvergenz:

- 
- (2) **Ordne jeden Datenpunkt seinem nächsten Clusterzentrum zu**
 - (3) **Aktualisiere jedes Clusterzentrum mit dem Mittelwert (Schwerpunkt) der zugeordneten Datenpunkte**

Schritt 1 : Initialisierung der Mittelwerte



TECHNISCHE
UNIVERSITÄT
DARMSTADT



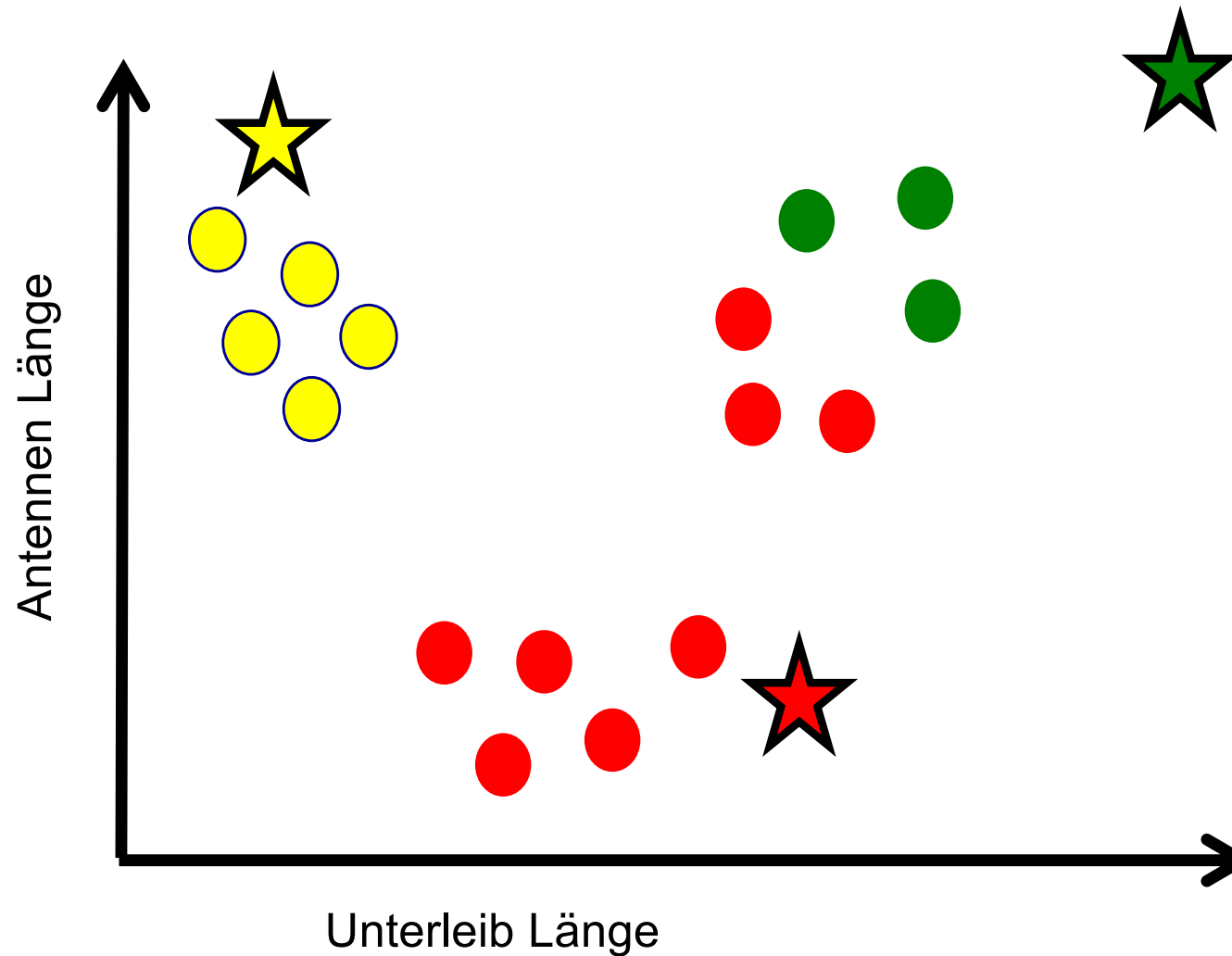
TECHNISCHE
UNIVERSITÄT
DARMSTADT



Schritt 2 : Zuordnung der Datenpunkte



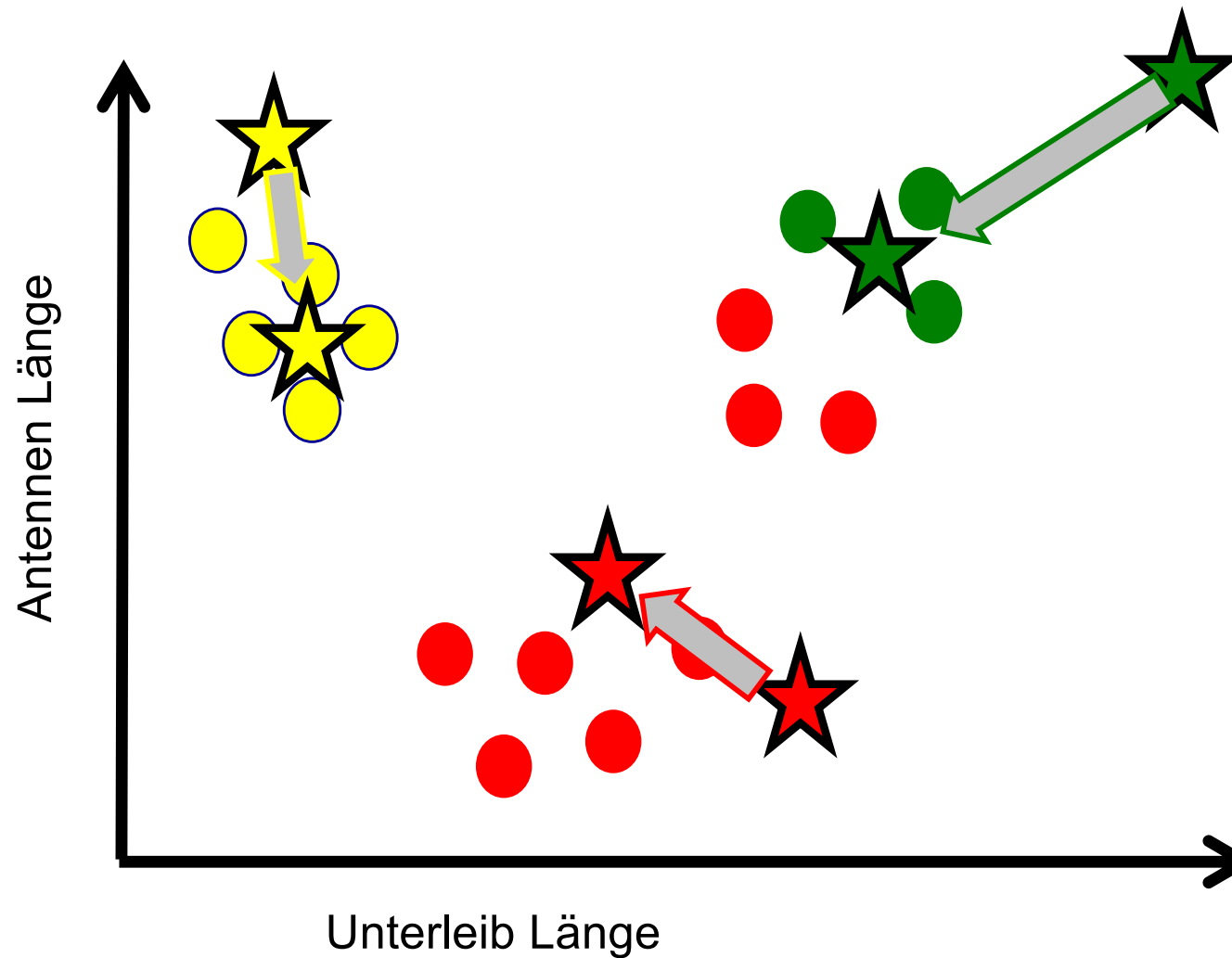
TECHNISCHE
UNIVERSITÄT
DARMSTADT



Schritt 3 : Aktualisierung der Clusterzentren



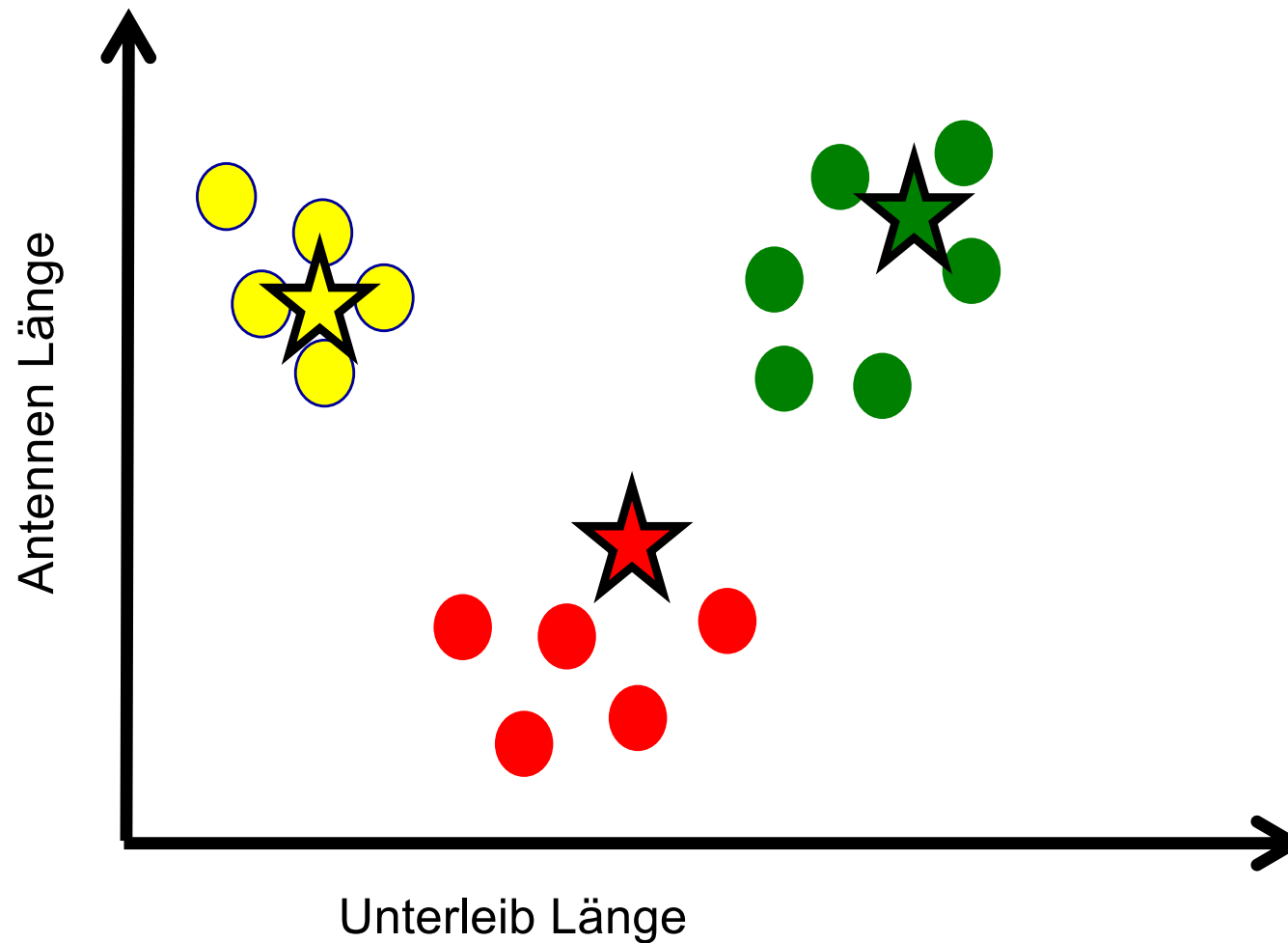
TECHNISCHE
UNIVERSITÄT
DARMSTADT



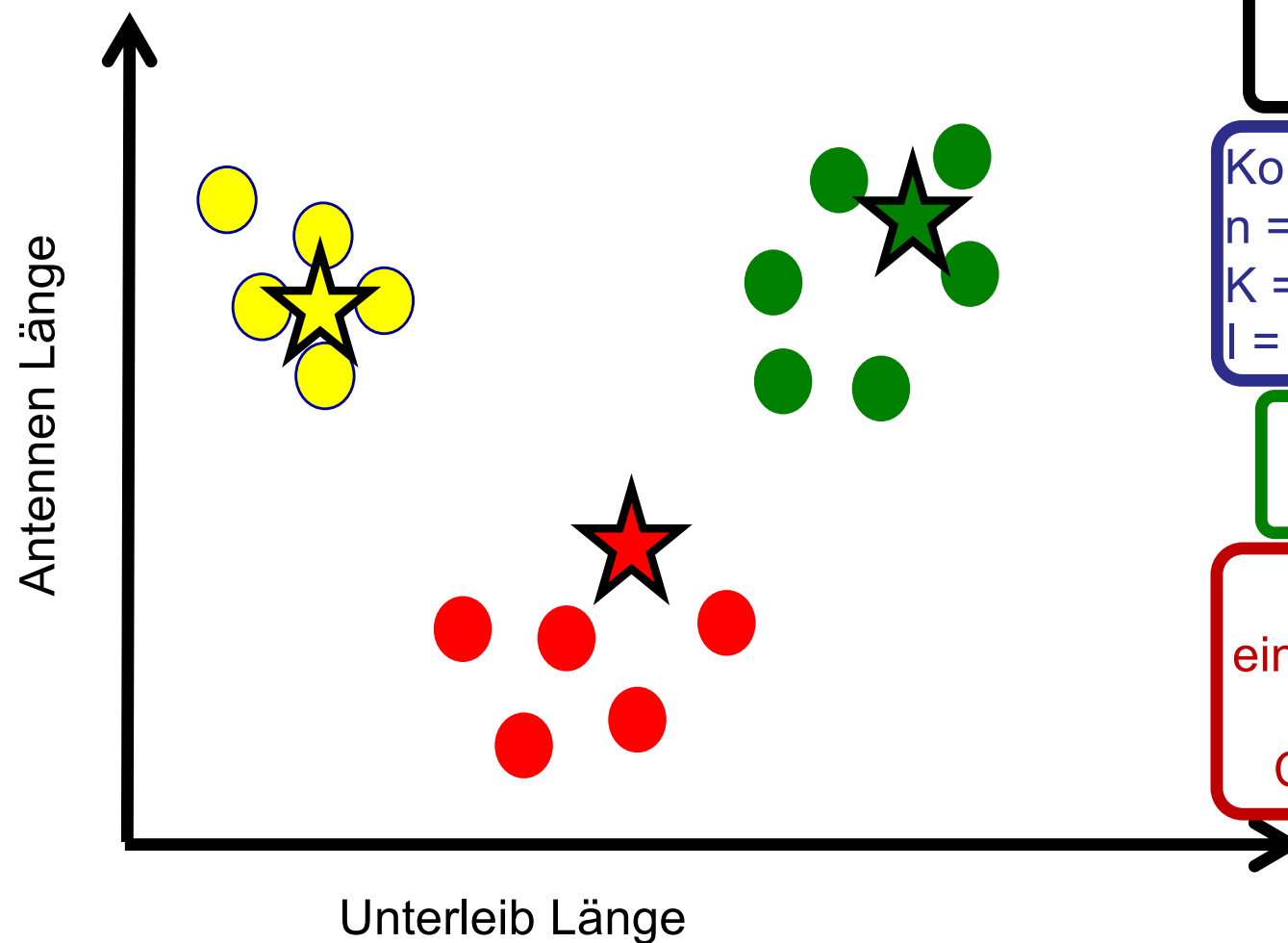
TECHNISCHE
UNIVERSITÄT
DARMSTADT



Schritt 2 : Zuordnung der Datenpunkte



Schritt 4: Algorithmus ist konvergiert



Clustering findet
Struktur in den Daten

Komplexität ist $O(n * K * I)$
 n = Zahl der Datenpunkte
 K = Zahl der Cluster
 I = Zahl der Iterationen

Güte hängt von der
Initialisierung ab

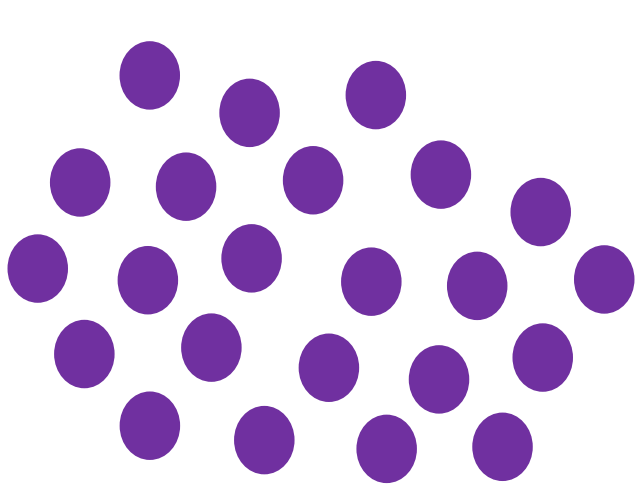
Ein gutes Clustering mit
einem kleinen K kann besser
sein als ein schlechtes
Clustering mit grossem K



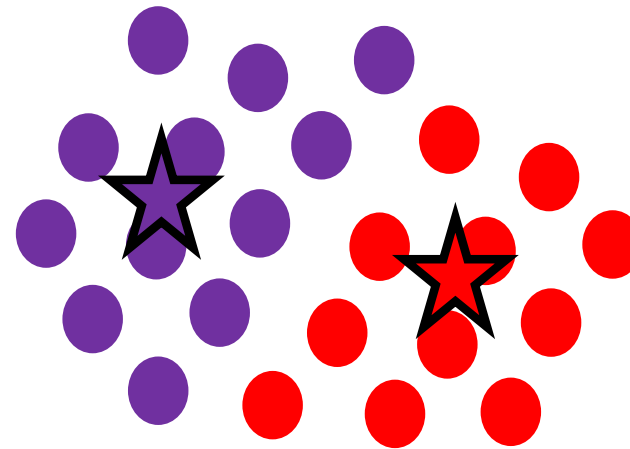
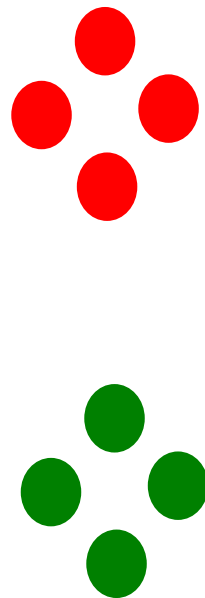
Nachteile des K-Means-Verfahrens

- Probleme bei Clustern mit unterschiedlicher Größe und Dichte
- Probleme mit nicht "kugel-förmigen" (spherical) Clustern
- Resultate hängen stark von der Anzahl und initialen Festlegung der Clusterzentren ab
- Wahl der Distanzmetrik hat großen Einfluss auf Cluster
- Ausreißer und leere Cluster führen zu Verzerrungen
- Fluch der hohen Dimension: In hochdimensionalen Räumen sind alle Daten unähnlich

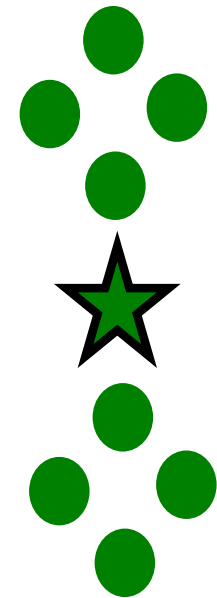
Nachteile: Unterschiedliche Größen



Daten



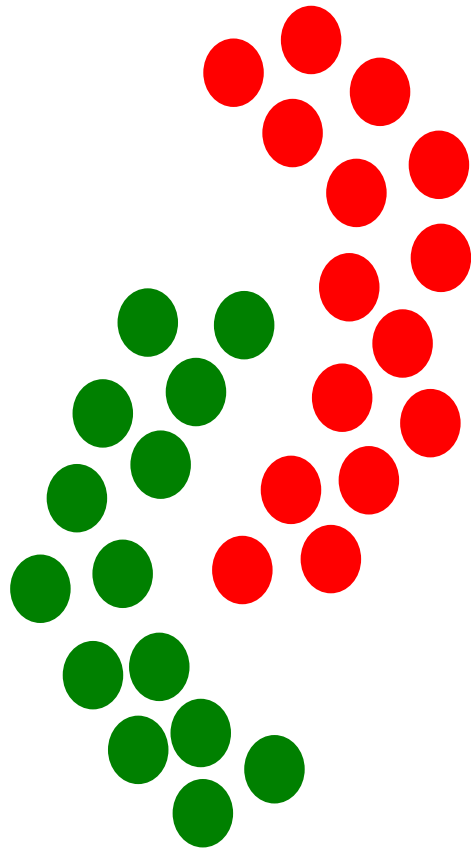
kMeans (3 Cluster)



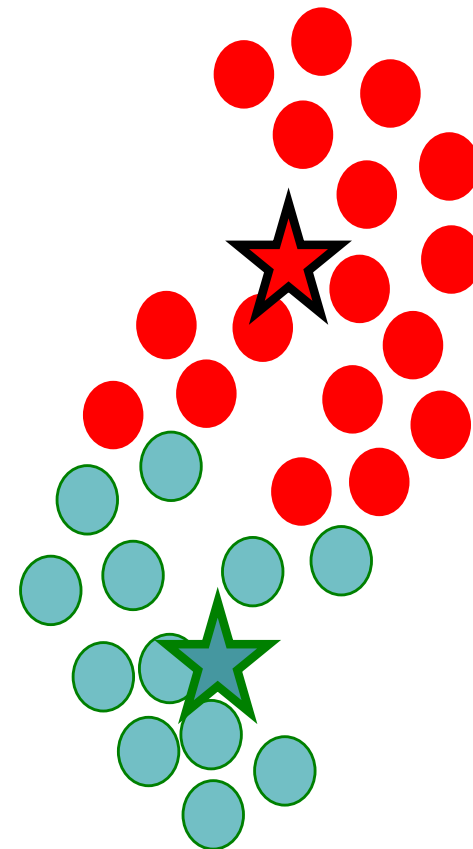
Nachteile: Nicht-kugelförmig



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Daten



kMeans (2 Cluster)



TECHNISCHE
UNIVERSITÄT
DARMSTADT





Strategien zum Umgehen der Nachteile

- Mehrfaches Durchführen + Behalten des besten Ergebnisses
- „Over-Clustering“ + Nachverarbeitung
- Probabilistische oder kernelized Varianten
- Andere Verfahren zur Clusteranalyse wie z.B. Spectral Clustering, Random Projections, Clusteranalyse mit Randbedingungen, DBSCAN, Bi-Clustering, LDA, ...
- ...



Was wissen wir jetzt?

Die Lernaufgabe Clustering kennen Sie

Wir haben zwei Klassen von Methoden gesehen:

- hierarchisches Clustering,
- k-Means.

Die Wahl des Abstandmaßes ist entscheidend für das Clustering.