Multimodale Analyse

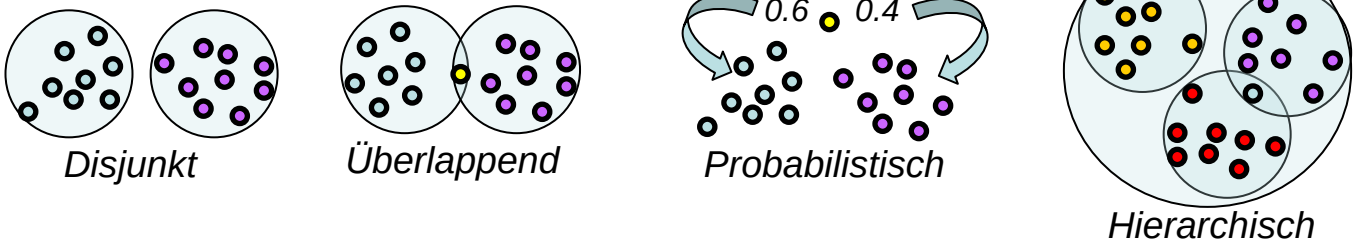
Clustering Verfahren

Elisabeth André
Chi Tai Dang
Stephan Hammer

Human Centered Multimedia
Institute of Computer Science
Augsburg University
Universitätsstraße. 6a
86159 Augsburg, Germany

Clustering Verfahren
Aufgabe und Zielsetzung

- Daten zu “natürlichen” Gruppen (**Clusters**) gruppieren
- Natürliche Gruppierung heißt:
  - Daten **innerhalb** einer Gruppe sind **ähnlich** zueinander
  - Daten verschiedener Gruppen sind unähnlich zueinander
- Vielzahl von Methoden zur Berechnung von Ähnlichkeit
- Vielzahl von Formen und Eigenschaften von Clustern
  - z.B. disjunkte vs. überlappende Cluster
  - z.B. deterministische vs. probabilistische Cluster
  - z.B. hierarchische vs. flache Cluster



Clustering Verfahren
Einsatzgebiete von Clustering

Universität Augsburg University

- Im Marketing zur Einordnung von Kunden in Gruppen und der Erstellung von gezielten Marketingstrategien
- Zur hierarchischen Strukturierung von Textdokumenten
- Erstellung von thematischen Karten aus Satellitenbildern
- In der Stadtplanung zur Gruppierung von Gebäudetypen
- Einordnung von Erdbeben-Epizentren zu Verwerfungen der Kontinentalplatten bei Untersuchung von Erdbeben
- Einordnung von Nutzern mit ähnlichem Interaktionsverhalten

Clustering Verfahren
Clustering vs. Klassifikation

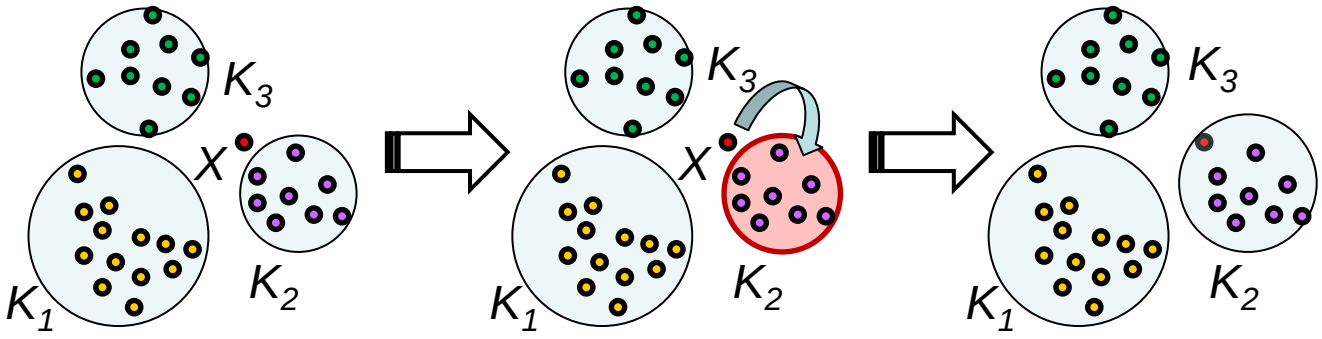
Universität Augsburg University

- Unterschiede zwischen **Clustering** und **Klassifikation**
  - Bei der Klassifikation sind die Klassen in die klassifiziert werden soll bereits vorgegeben.
  - Bei Clustering werden die Klassen zunächst noch durch eine bestimmte **Cluster-Analyse** gesucht.

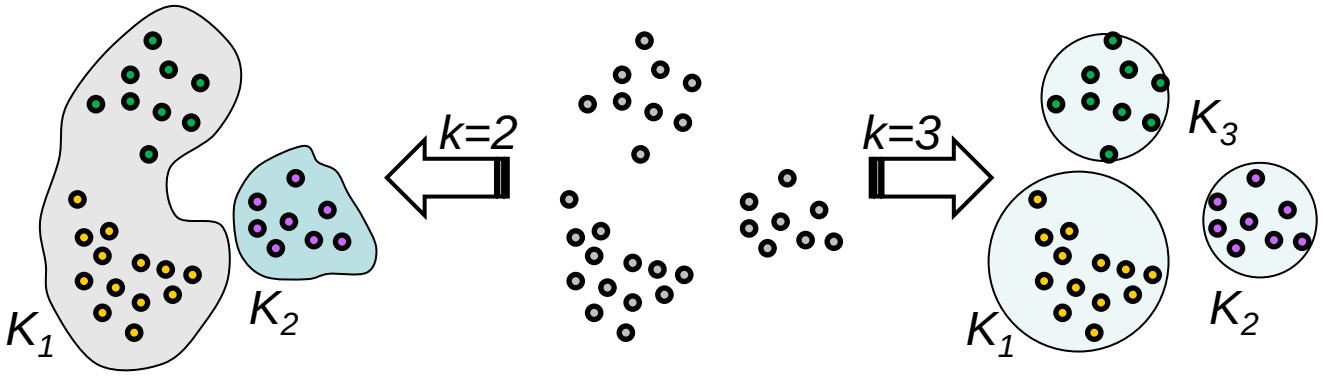
→ Clustering ein **nicht überwachtes Lernverfahren**

Clustering Verfahren
Cluster-Zuordnung

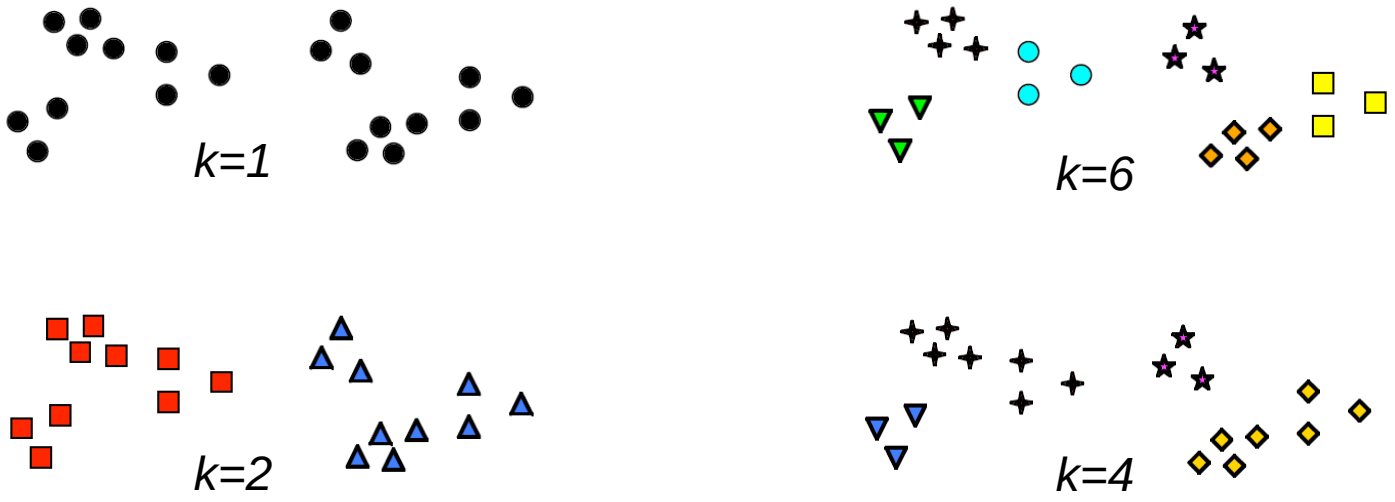
- **Gegeben:**
  - Eine bereits vorgegebene Menge von *k* Clustern K_1 bis K_k
  - Ein neues Objekt *X* das noch keinem der Cluster angehört
- **Gesucht:**
  - Zuordnung des Objekts *X* zu einem der *k* Cluster K_1 bis K_k
- **Ansatz:**
  1. Vergleiche das Objekt *X* zu allen Instanzen jedes Clusters.
  2. Wähle das Cluster, dessen Objekte *X* am ähnlichsten sind.



- **Gegeben:**
  - Datensatz gegeben als eine Menge von *n* Objekten X_1 bis X_n
- **Gesucht**
  - Diejenigen *k* Klassen K_1 bis K_k welche die *n* Objekte gemäß maximaler Ähnlichkeit gruppieren. Voraussetzungen dabei:
    - Die Anzahl der Klassen *k* ist bei Beginn **bereits vorgegeben**
    - Die Anzahl der Klassen *k* hängt ab von einer vorgegebenen **Mindestdistanz**, die von den Objekten die in verschiedenen Klassen liegen eingehalten werden muss.



- Cluster-Analyse ist **nicht eindeutig** und hängt stark von der **Anzahl** der Cluster und den **Ähnlichkeitsmaßen** ab



Clustering Verfahren
Ansätze zum Clustering

Universität Augsburg University

Verfahren zur Clusterbildung

Bottom-Up

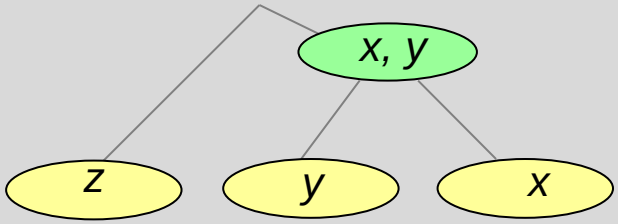
Beginne mit vielen Klassen und verschmelze diese in immer größer werdende Klassen.

Top-Down

Beginne mit einer Klasse und teile diese immer weiter auf. (wird kaum praktiziert)

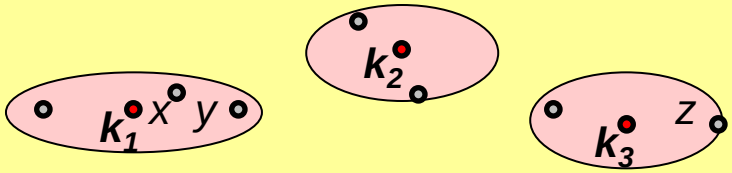
Hierarchisch-Agglomerative Verfahren

Hierarchie entsteht durch die Zusammenlegung von elementaren Klassen



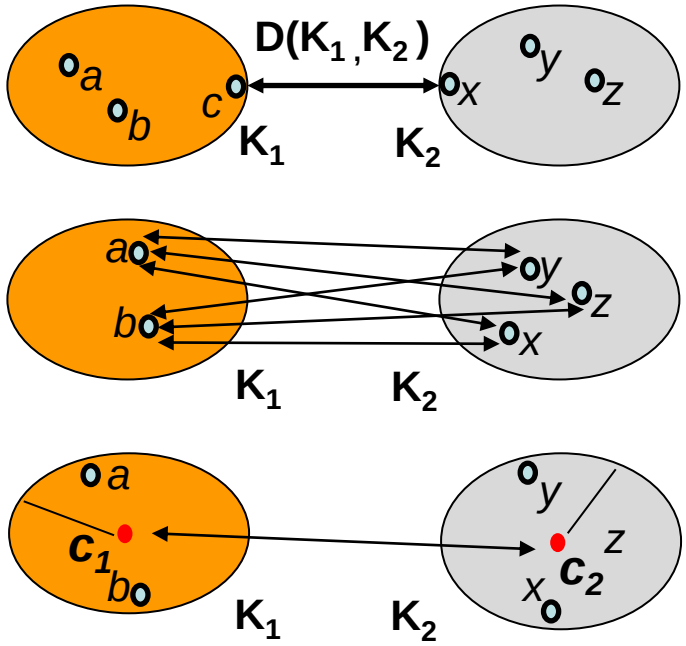
Partitionierende Verfahren

Klassenzahl *k* wird zusammen mit *k* Kristallisationskernen vorgegeben. Neue Objekte werden gemäß ihres Abstands zum nächstgelegenen Kristallisationspunkt eingeordnet. Mit wachsenden Klassen, können Kristallisationskerne neu festgelegt werden.



Clustering Verfahren
Methoden der Distanzmessung

- Die Bestimmung der **Distanz** zwischen zwei Clustern ist das Kriterium zur Verschmelzung der beiden Cluster oder um zu entscheiden welcher Cluster einem neuen Objekt am **nächsten** liegt:
  - **Minimum/Maximum:** Distanz des nächsten bzw. weitesten Nachbarn.
  - **Gruppenmittel**: Mittlerer paarweiser Abstand zwischen Klasseninstanzen.
  - **Centroid-Abstand**: Abstände der Mittelpunkte der einzelnen Cluster.



Clustering Verfahren
Verschiedene Distanzmaße

Universität Augsburg University

- **Euklidische Norm:**

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2}$$

- **Manhattan Norm:**

$$d(p,q) = |(p_1 - q_1)| + |(p_2 - q_2)| + \ldots + |(p_n - q_n)|$$

$$p = (p_1, \ldots, p_n)$$
$$q = (q_1, \ldots, q_n)$$

- **Maximum Norm:**

$$d(p,q) = max\{|p_1 - q_1|, |p_2 - q_2|, \ldots, |p_n - q_n|\}$$

- **Minkowski Norm:**

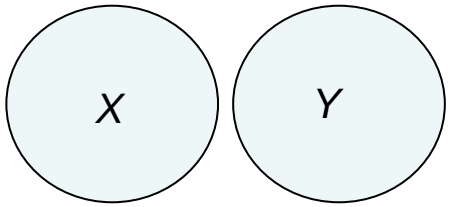
$$d(p,q) = \left(|(p_1 - q_1)|^x + |(p_2 - q_2)|^x + \ldots + |(p_n - q_n)|^x\right)^{\frac{1}{x}}$$

- Die vorherigen Abstandsmaße werden in der Regel für **numerische** Werte eingesetzt. Bei Daten von Objekten $X=(x_1,\ldots, x_n)$ mit **nominalen** Attributwerten x_i, wie z.B. *„Geschlecht“* verwendet man als Abstandsmaß z.B.

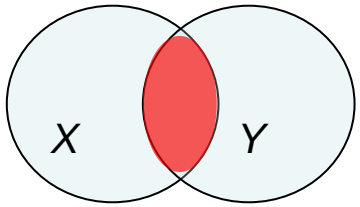
$$d(X,Y)=\sum_{i=1}^{n} d(x_i,y_i), \quad d(x_i,y_i)=\begin{cases} 0, & wenn \quad x_i = y_i \\ 1, & wenn \quad x_i \neq y_i \end{cases}$$

- Für **endliche Mengen** als Objekte vergleicht man die Anzahl der verschiedenen Elemente innerhalb *X* und *Y*

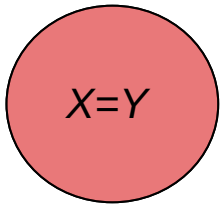
$$d(X,Y)=\frac{|X \cup Y| - |X \cap Y|}{|X \cup Y|}$$



d(X,Y)=1



0 < d(X,Y) < 1



d(X,Y) =0

Clustering Verfahren
Der K-Means Algorithmus

- **Initialisierung** (Produziere zunächst *k* disjunkte Cluster)
  - Wähle für jeden der *k* Cluster einen Clusterschwerpunkt S_i
  - S_i kann dabei bereits ein willkürliches Objekt X_i sein oder man wählt S_i nach Belieben, d.h. noch keines der Objekte $X_1 \ldots X_n$
- **Einordnung** (Für alle einzuordnenden Objekte $X_1 \ldots X_n$)
  - Errechne den Abstand von X_i zu jedem der *k* Schwerpunkte
  - Lege X_i in den Cluster mit geringstem Schwerpunkt-Abstand
- **Berechnung** (Für alle Cluster $K_1 \ldots K_k$)
  - Berechne den Schwerpunkt für K_i neu denn durch Zuordnung der Objekte können sich Cluster-Schwerpunkte verschieben.
- **Wiederholung** (Bis alle Clusterschwerpunkte S_i stabil sind)
  - Wiederhole die Schritte Einordnung und Berechnung

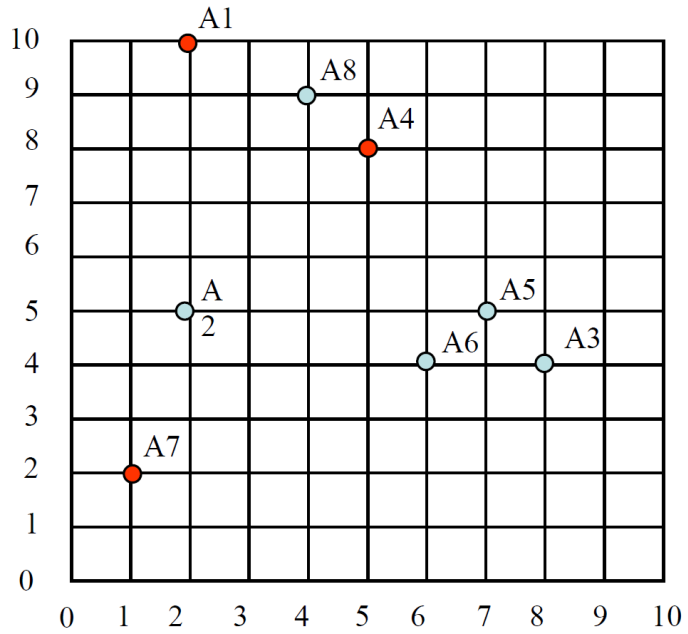
Clustering Verfahren
K-Means Beispielaufgabe

Universität Augsburg University

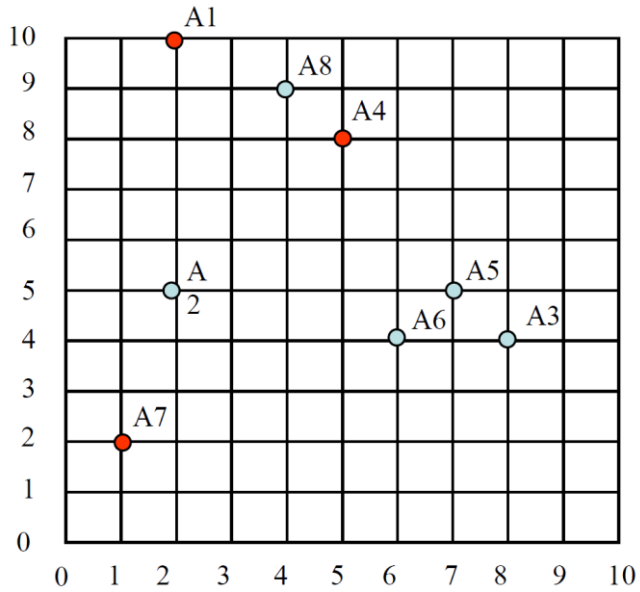
- *8* Datensätze $A_1,\ldots, A_8$ sollen auf *3* Cluster aufteilt werden, d.h. $k=3$
- Als das Distanzmaß soll die *Euklidsche Distanz* verwendet werden
- Als Schwerpunkte für die Cluster wähle $S_1=A_1$, $S_2=A_4$ und $S_3=A_7$

1. Trage die Punkte in ein 2-dimensionales **Koordinatensystem** ein

$$A_1 = (2,10)$$
$$A_2 = (2,5)$$
$$A_3 = (8,4)$$
$$A_4 = (5,8)$$
$$A_5 = (7,5)$$
$$A_6 = (6,4)$$
$$A_7 = (1,2)$$
$$A_8 = (4,9)$$



2. Berechne **Distanzmatrix** mit den Distanzen zwischen den Punkten

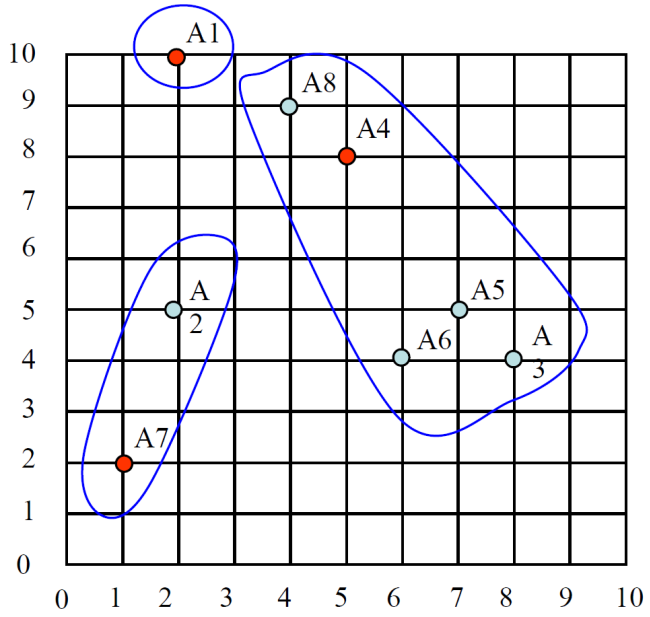


	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{72}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Clustering Verfahren
K-Means Beispielaufgabe

Universität Augsburg University

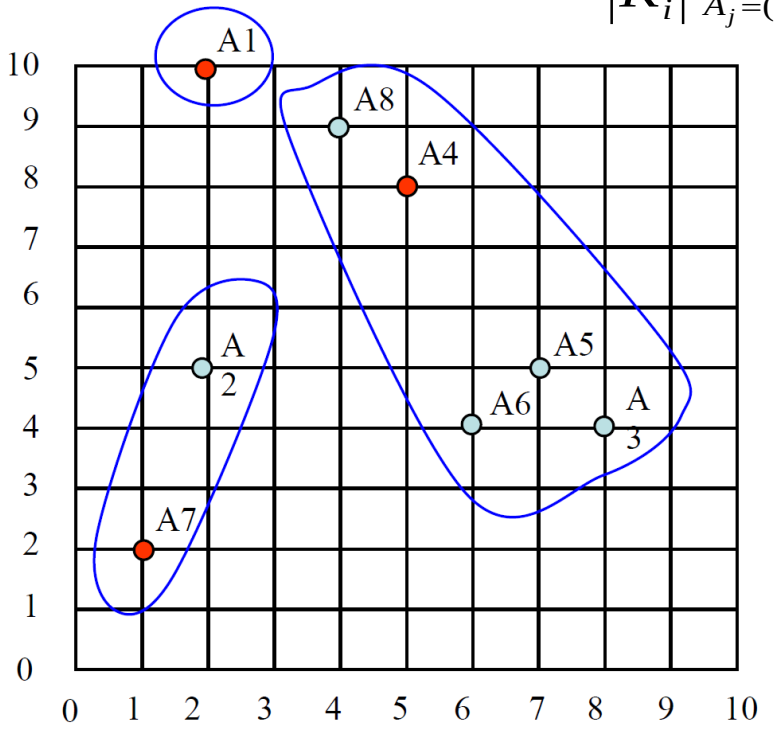
3. Ordne die Punkte $A_1,\ldots, A_8$ gemäß ihrem Abstand zu den drei Schwerpunkten S_1, S_2 und S_3 in einen der bisherigen Cluster ein

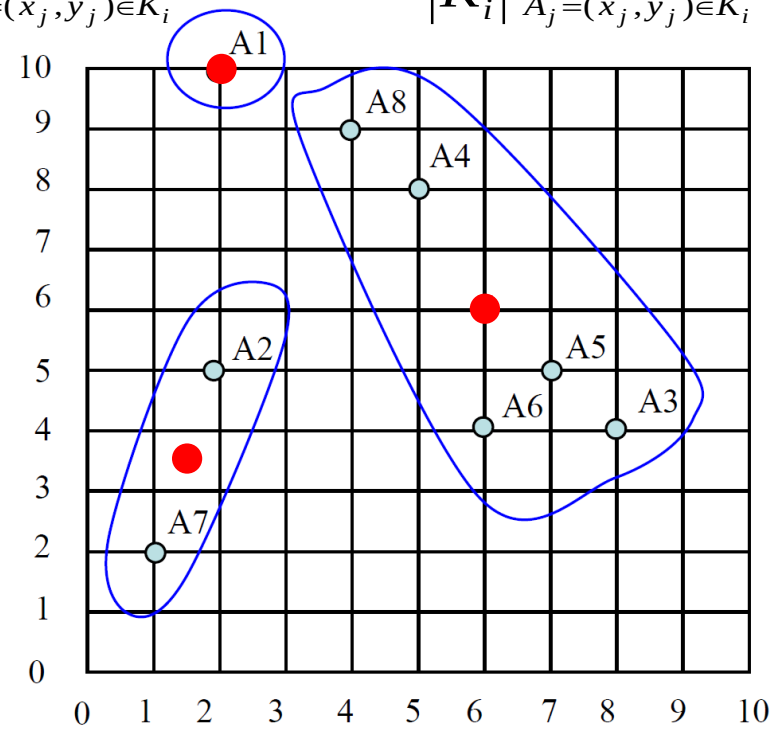


	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{72}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2	$\sqrt{25}$	0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3	$\sqrt{36}$	$\sqrt{37}$	0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4	$\sqrt{13}$	$\sqrt{18}$	$\sqrt{25}$	0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5	$\sqrt{50}$	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{13}$	0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6	$\sqrt{52}$	$\sqrt{17}$	$\sqrt{2}$	$\sqrt{17}$	$\sqrt{2}$	0	$\sqrt{29}$	$\sqrt{29}$
A7	$\sqrt{65}$	$\sqrt{10}$	$\sqrt{53}$	$\sqrt{52}$	$\sqrt{45}$	$\sqrt{29}$	0	$\sqrt{58}$
A8	$\sqrt{5}$	$\sqrt{20}$	$\sqrt{41}$	$\sqrt{2}$	$\sqrt{25}$	$\sqrt{29}$	$\sqrt{58}$	0

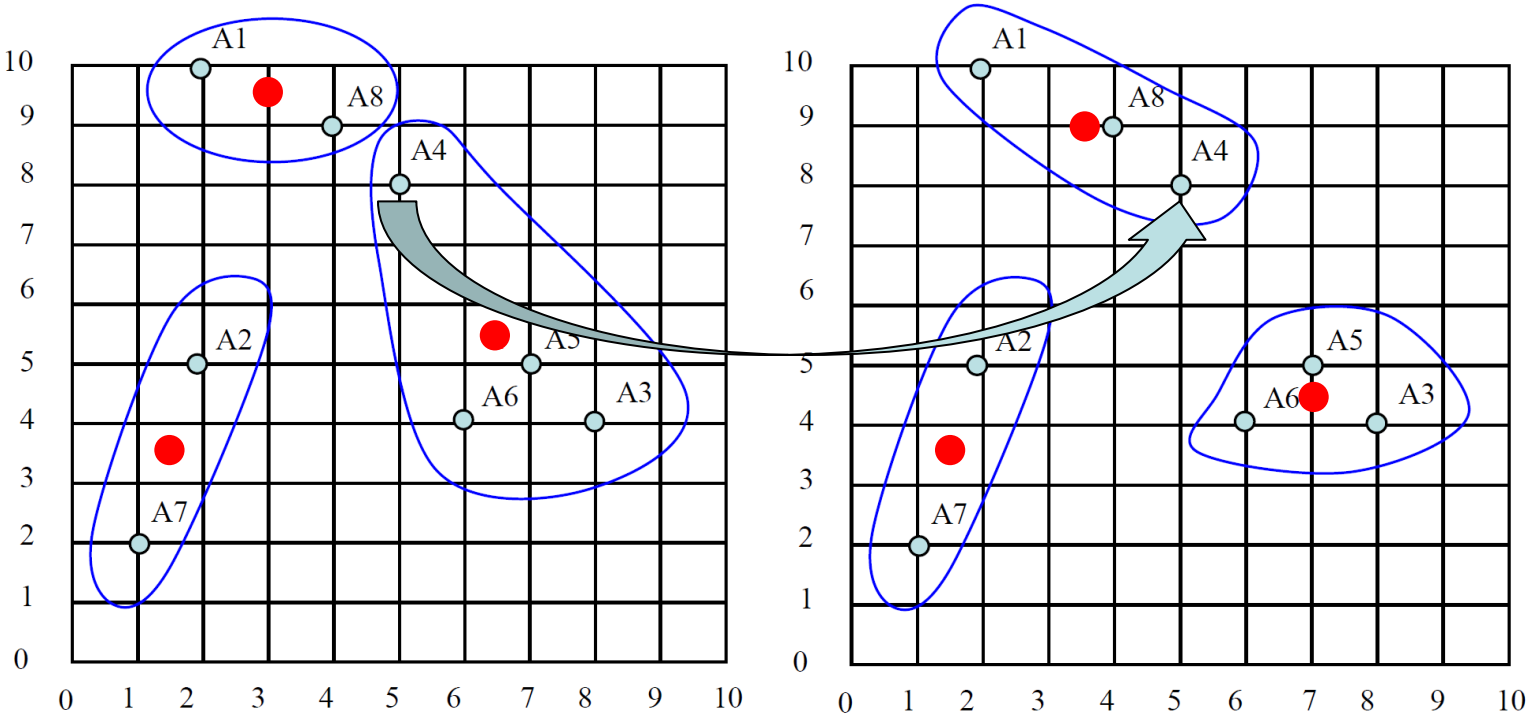
4. Bestimme für jeden Cluster einen neuen Schwerpunkt S_1, S_2 und S_3

$$S_i = (X_i, Y_i) \quad mit \quad X_i = \frac{1}{|K_i|} \sum_{A_j=(x_j,y_j)\in K_i} x_i \quad und \; Y_i = \frac{1}{|K_i|} \sum_{A_j=(x_j,y_j)\in K_i} y_i$$





5. Berechne ausgehend von den neuen Schwerpunkten S_1,S_2 und S_3 die **neue Distanzmatrix** mit den Abständen aller Punkte zu diesen neuen Schwerpunkten und **ordne** die Punkte **neu in** die **Cluster**.



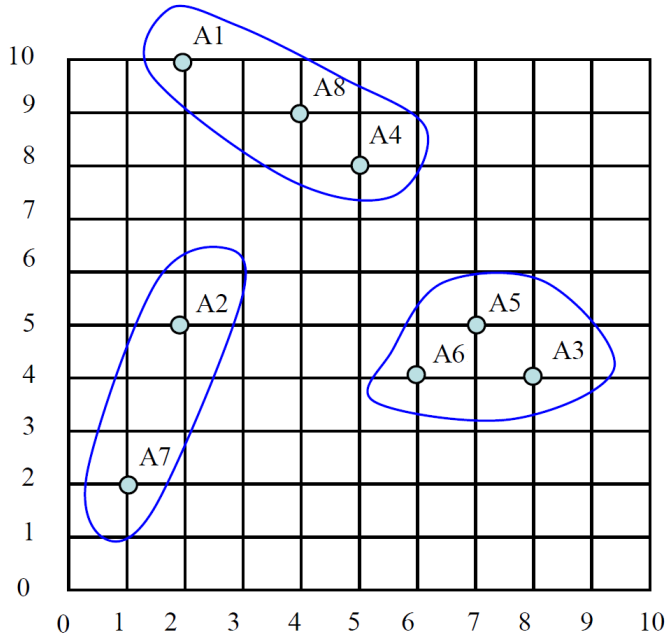
nach 2 Iterationen

nach 3 Iterationen

Clustering Verfahren
K-Means Beispielaufgabe

Universität Augsburg University

6. Wiederhole Schritt *5* solange bis sich die Cluster nicht mehr ändern



nach 4 Iterationen

7. Wann immer ein neues Objekt hinzukommt führe Neuberechnung der Clusterzuordnung und der neuen Cluster-Schwerpunkte durch

Clustering Verfahren
K-Means Eigenschaften

Universität Augsburg University

- Je nach Verfahren zur Abstandsmessung kann sich der Abstand durch die Hinzunahme eines weiteren Objekts ändern (z.B. bei der Schwerpunktsmethode oder der des nächsten Nachbarn) → Daher sind eventuell mehrere Iterationen notwendig, bis stabile Cluster vorliegen.
- Wählt man für die initialen Schwerpunkte S_j jeweils die Objekte X_i, so enthält jeder Cluster immer mindestens ein Element und dann überlappen die Cluster nicht und sind nicht hierarchisch.
- Die Cluster-Initialisierung hat starken Einfluss auf das Endergebnis und die Anzahl der benötigten Iterationen.

Clustering Verfahren
Der K-Nearest Neighbor (kNN)

Universität Augsburg University

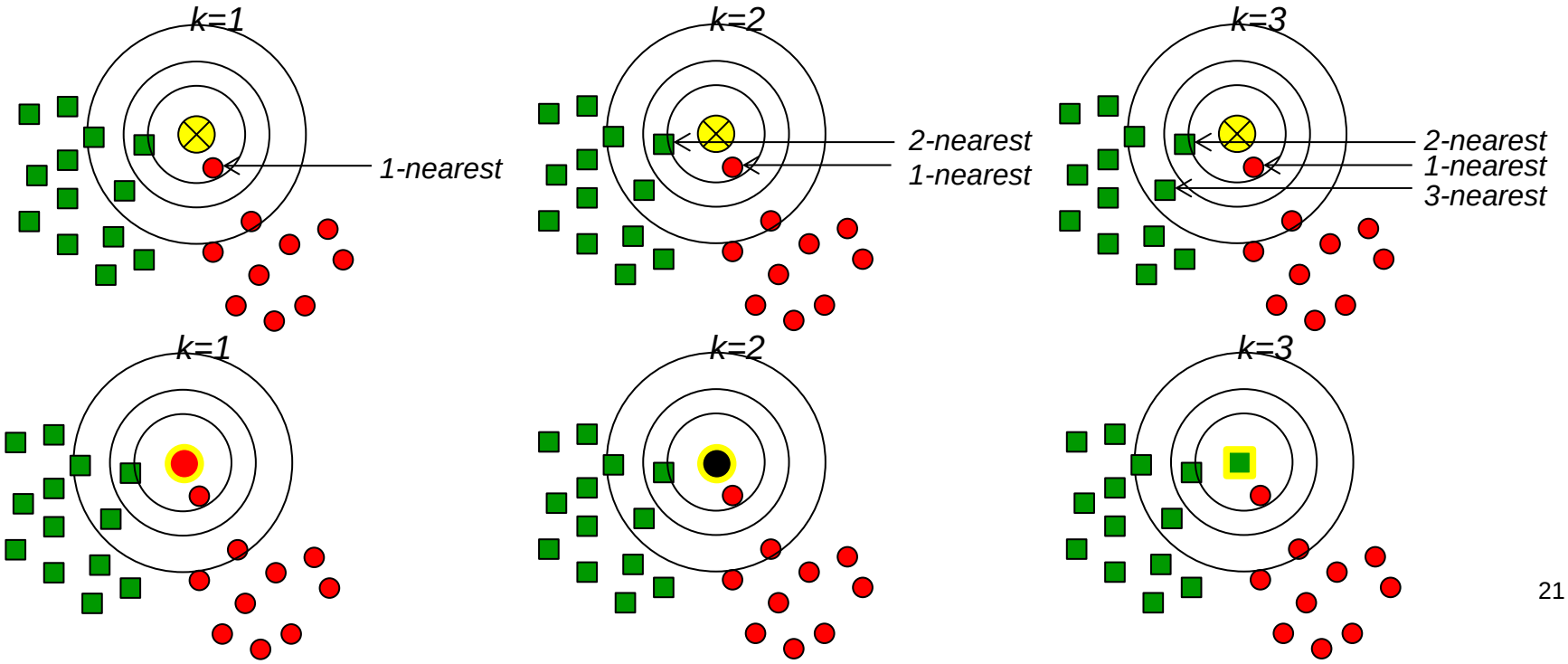
- **Gegeben:**
  - Eine Menge von Beispielobjekten $B_1,...,B_n$
  - Eine neue einzuordnende Objekt Instanz X
- **Gesucht:**
  - Das Cluster mit den k nächsten Nachbarpunkten von X

    (Im Gegensatz zu Cluster mit nächstem Schwerpunkt)
- **Ansatz:**
  1. Berechne den Abstand $D(X, B_i)$ für alle Beispielobjekte B_i
  2. Seien dann $B_1(X),...,B_k(X)$ die k nächsten Nachbarn von X, dann bestimme nun das Cluster K, dem die meisten der $B_i(X)$ angehören und ordne dann X diesem Cluster K zu.

Clustering Verfahren
Der K-Nearest Neighbor (kNN)

Universität Augsburg University

- Oft erhält man **unterschiedliche Cluster** für verschiedene Werte von *k* → Wie soll man *k* wählen?
- Meist untersucht man einen Bereich von Werten für *k* und wählt dann dasjenige *k* mit den besten Ergebnissen



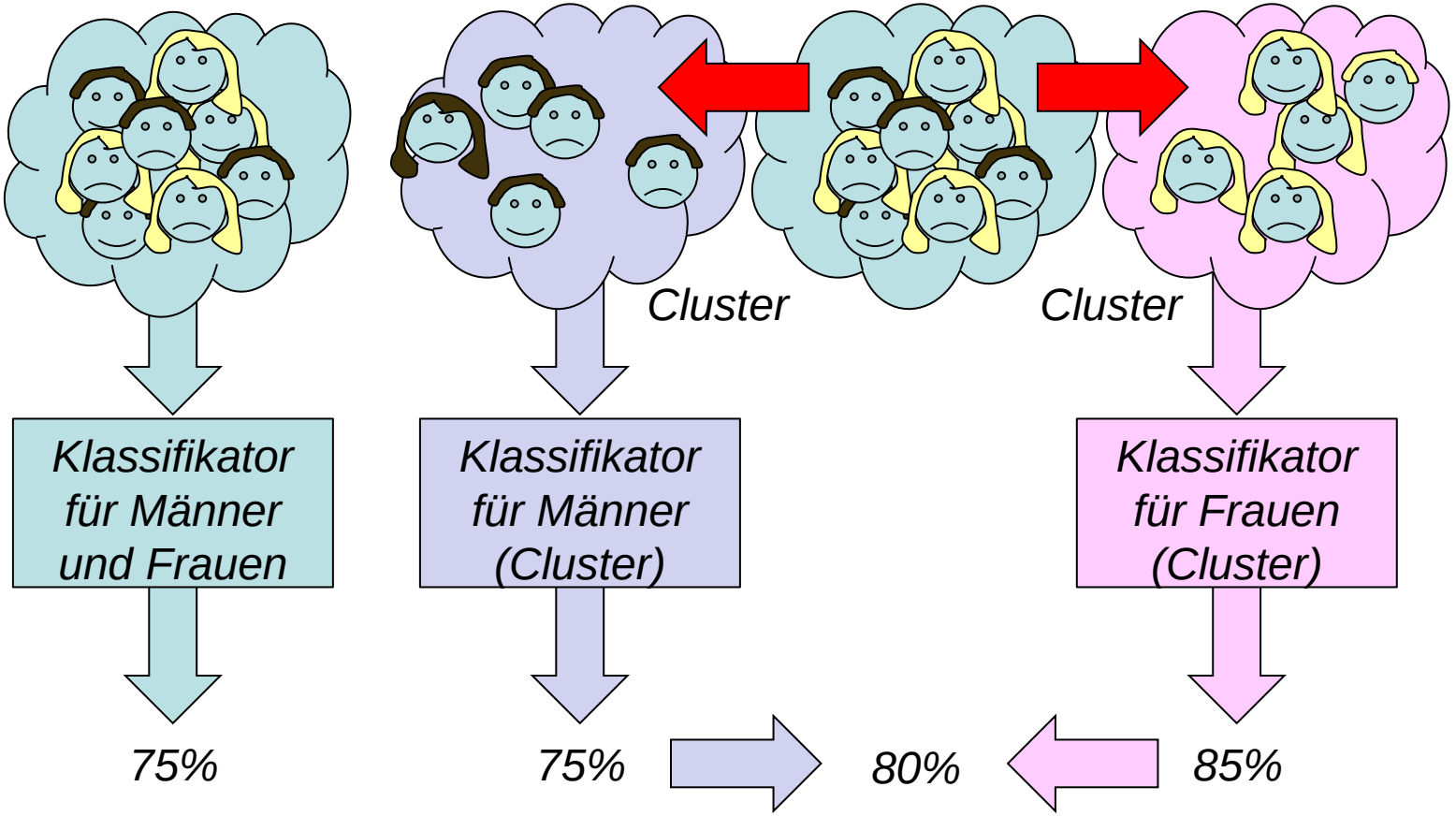
- **Problem:**
  - Wie evaluiert man die Performanz eines Cluster Ergebnis?
  - Woher wissen wir, dass die Cluster korrekt bzw. gut sind?
- **Lösung:**
  - **Intuitiv:** Teste die Cluster intuitiv. Überprüfe, ob sie Sinn machen, d.h. nach eigenem Ermessen sinnvoll einordnen.
  - **Experte:** Lasse einen Experten Cluster manuell erstellen und vergleiche sie anschließend mit den automatisch Generierten.
  - **Vergleich:** Vergleiche die Cluster mit einer vordefinierten Klassifikation, sofern es eine gibt und sie bekanntlich gut ist
  - **Aufgabe:** Führe eine aufgabenbasierte Evaluation durch, d.h. überprüfe, ob ein anderer Algorithmus durch die Verwendung der Cluster verbessert werden kann.

Clustering Verfahren Evaluation der Ergebnisse

Universität Augsburg University

- **Beispiel:**
  - Clusteralgorithmus zur Gruppierung von Audioaufnahmen
- **Intuitiv oder Experte:**
  - Überprüfe selbst wie sich die Audioaufnahmen in Cluster charakterisieren lassen oder frage einfach Dieter Bohlen.
- **Vergleich:**
  - Schaue nach ob die Cluster mit den Ergebnissen verschiedener Klassifikatoren (z.B. Geschlecht, Alter, etc.) korrespondieren.
- **Aufgabe:**
  - Prüfe, ob sich durch das Clustering die Emotionserkennung verbessern lässt. Vergleiche Ergebnis eines Klassifikators für alle Daten mit zwei anderen Klassifikatoren, die auf ein Merkmal trainiert wurden.

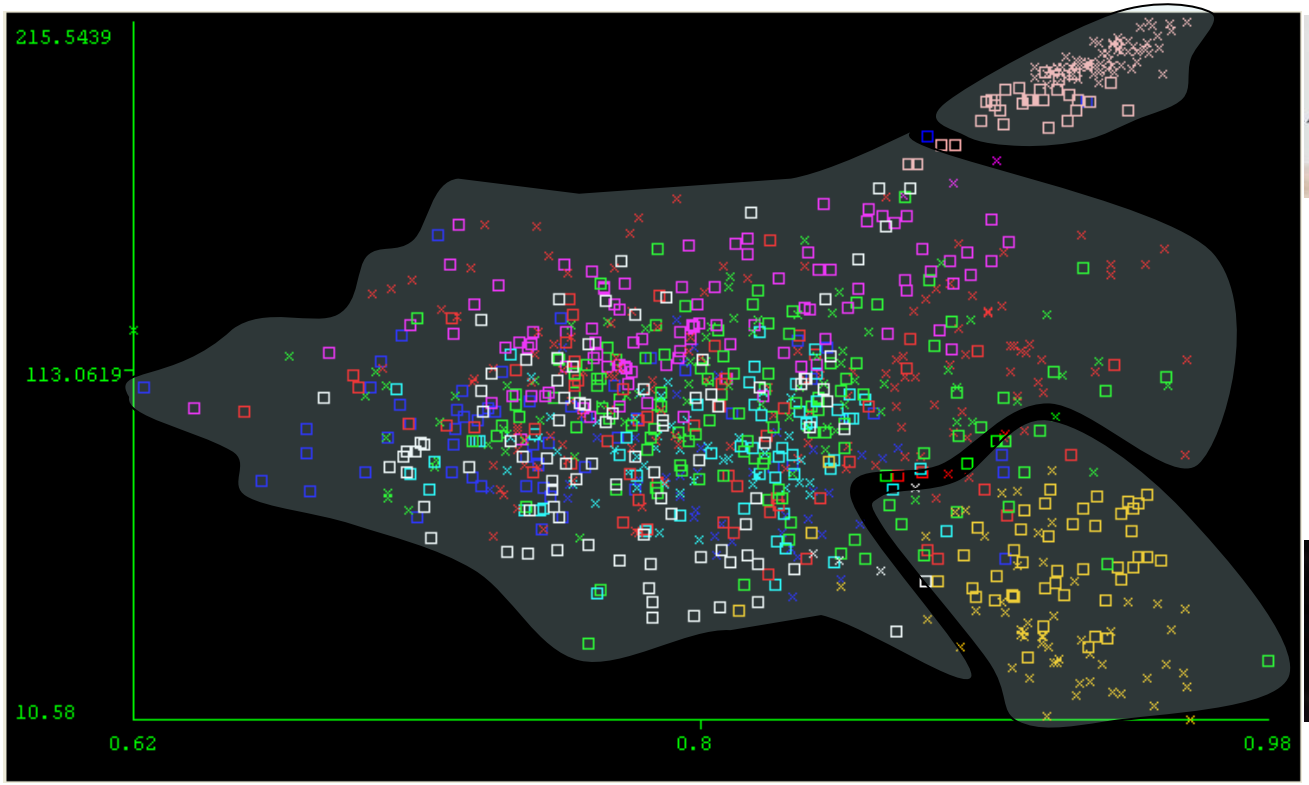
- **Beispiel:**
  - Wird durch das Clustering die Emotionserkennung besser?



Clustering Verfahren
Evaluation der Ergebnisse

Universität Augsburg University

- Manuelles intuitives Überprüfen der Cluster Ergebnisse:



Ansonsten total chaotisch überlappende Cluster

Hauptsächlich Zeichnungen von Dinosauriern und Vögeln

Hauptsächlich Fotographien von Blüten und Blumen

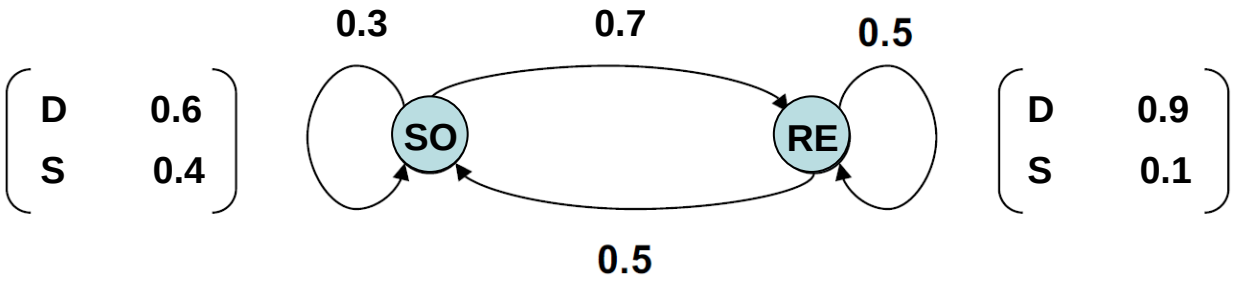
Hidden Markov Modelle
Beispiel für Training

Universität Augsburg University

- Gefangener im Verlies

Ein Gefangener im Kerkerverlies möchte das aktuelle Wetter herausfinden. Er schätzt, dass die Schuhe der Wärter bei Regen zu 90 % dreckig, bei sonnigem Wetter aber nur zu 60 % dreckig sind, so kann er durch Beobachtung der Wärterschuhe Rückschlüsse über das Wetter ziehen. Zu Beginn geht er davon aus, dass alle Wetteränderungen gleichwahrscheinlich sind.

Ergebnis nach Training (für O = Sauber, Dreckig, Dreckig):

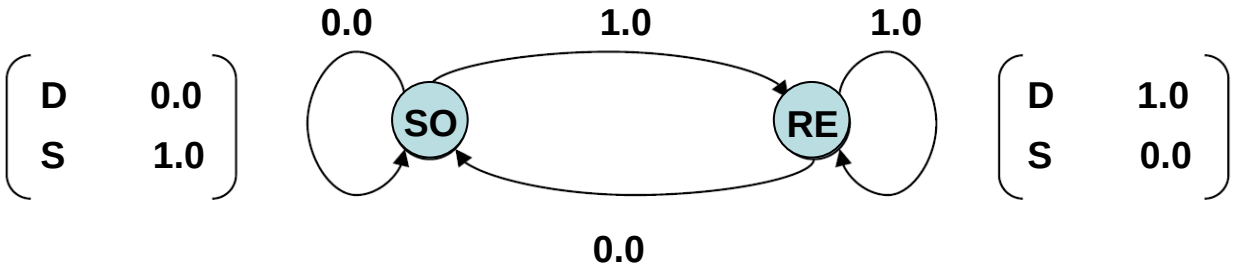


Hidden Markov Modelle
Wahl der Wahrscheinlichkeiten

Universität Augsburg University

- Auch wenn man wenig Informationen hat, müssen die Wahrscheinlich-keiten vernünftig gewählt werden.
- Eine unbedachte Wahl kann ein Training des Modells unmöglich machen.

- Beispiel:



Clustering

Bonusaufgabe

Clustering Verfahren Bonusaufgabe

- Sie haben verschiedene Clustering-Verfahren sowie Methoden und Maße zur Berechnung der Distanz von Merkmalen kennengelernt.

- Sie sollen Clustering anwenden um **einzelne Hunde** basierend auf ihren **äußeren Merkmalen** in Cluster einzuordnen. Die Cluster sollen später im Idealfall die verschiedenen **Hunderassen** repräsentieren.

Mira *(Brandlbracke)*

Größe: 52cm
Gewicht: 24kg
Farbe: Dunkel
Fell: Kurz

Kurt *(Kurzhaardackel)*

Größe : 26cm
Gewicht: 8kg
Farbe: Dunkel
Fell: Kurz

Paul *(Jack Russel)*

Größe : 25cm
Gewicht: 7kg
Farbe: Hell
Fell: Kurz

Tom *(Jack Russel)*

Größe : 23cm
Gewicht: 5kg
Farbe: Hell
Fell: Mittel

Clustering Verfahren Bonusaufgabe

Universität Augsburg University

1. Finden Sie **ein** Merkmal (Größe, Gewicht, Farbe, Fell) das sehr gut geeignet ist um ein klare Trennung beim Clustering vorzunehmen?

2. Finden Sie eine sehr gut geeignete **Kombination** von Merkmalen?

Mira *(Brandlbracke)*

Größe: 52cm
Gewicht: 24kg
Farbe: Dunkel
Fell: Kurz

Kurt *(Kurzhaardackel)*

Größe : 26cm
Gewicht: 8kg
Farbe: Dunkel
Fell: Kurz

Paul *(Jack Russel)*

Größe : 25cm
Gewicht: 7kg
Farbe: Hell
Fell: Kurz

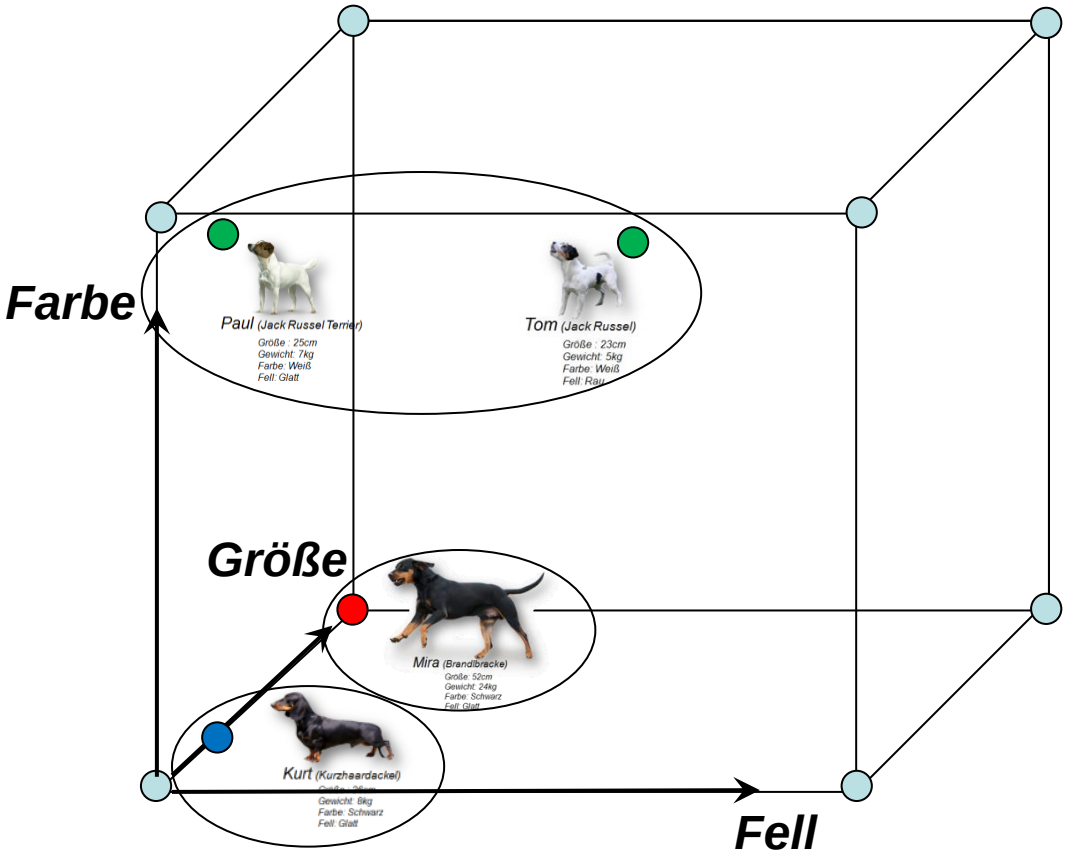
Tom *(Jack Russel)*

Größe : 23cm
Gewicht: 5kg
Farbe: Hell
Fell: Mittel

Clustering Verfahren Bonusaufgabe

Universität Augsburg University

1. **Keines** der Merkmale ist **alleine geeignet** für eine klare Trennung!
2. Die Kombination aus **Größe** und **Farbe** liefert gute Abgrenzungen!



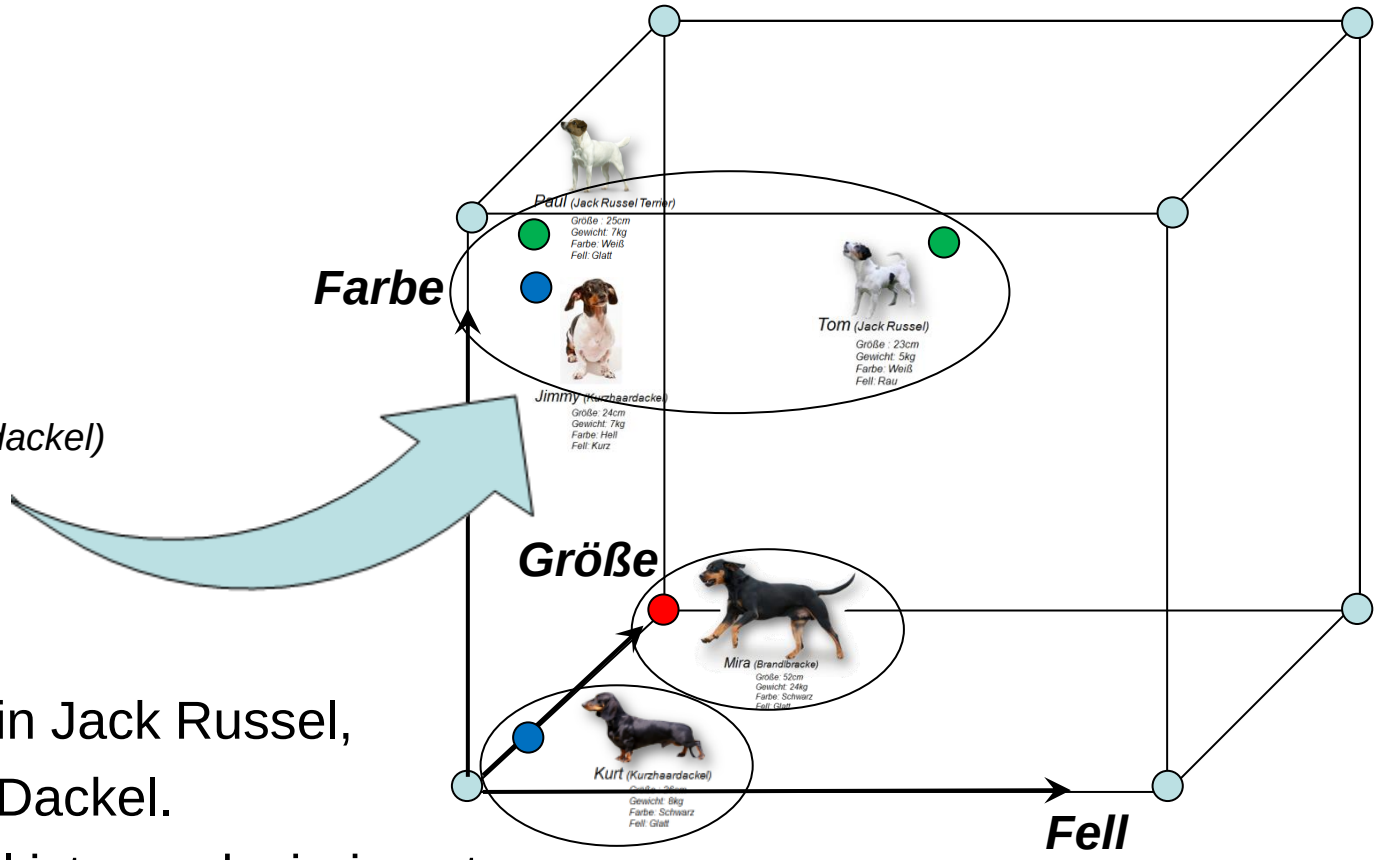
Clustering Verfahren Bonusaufgabe

Universität Augsburg University

- Aber **Achtung**! Was passiert, wenn jetzt noch Jimmy dazu kommt?

Jimmy *(Kurzhaardackel)*

Größe: 24cm
Gewicht: 7kg
Farbe: Hell
Fell: Kurz



- Jimmy ist kein Jack Russel, sondern ein Dackel.
  → Manchmal ist es schwierig gute Abgrenzungs-Merkmale zu finden.