Seminar Softwareentwicklung in der Wissenschaft

Unsupervised Deep Learning

Betreuer: Jakob Luettgau, Tobias Finn

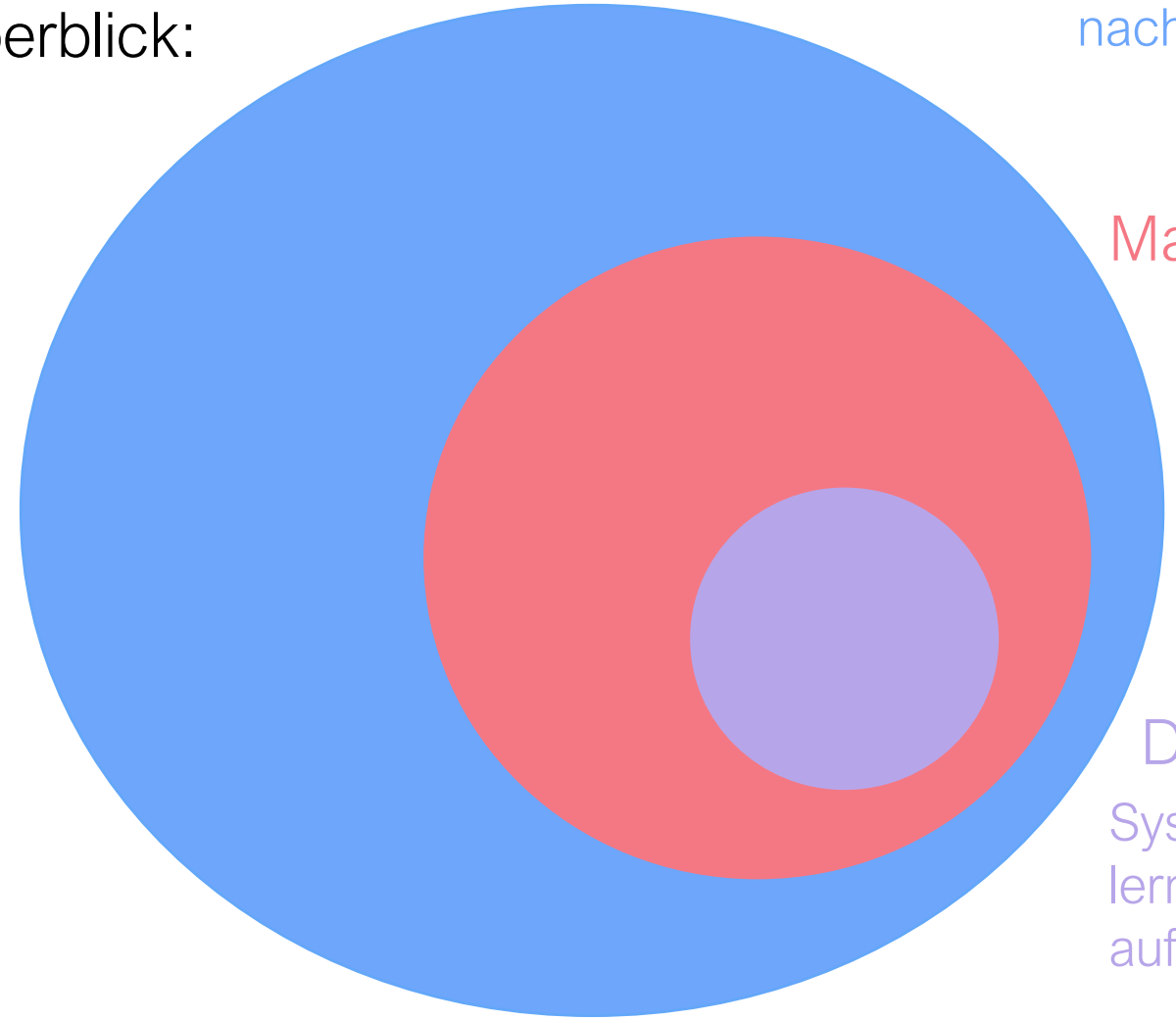
von Jennifer Heizenreider

Hamburg, den 21. Juni 2021

Gliederung

Definitionen

Überblick:



Zusammenhang zwischen ML, DL und KI

Künstliche Intelligenz

Systeme, die menschliches Denken und Handeln nachahmen.

Machine Learning

Systeme, die mit einem Algorithmus aus Erfahrungen (Daten) lernen. Der Algorithmus verbessert sich, durch menschliches Feedback.

Deep Learning

Systeme, die ohne menschliche Anleitung lernen. Das System prüft selbst, ob sich aufgrund eines Input etwas verändert.

Supervised Machine Learning

„Überwachtes Lernen“

Beispiel: Algorithmen trainieren Orchideen von Knabenkraut zu unterscheiden.

Datenset→ Input = Bilder

Knabenkraut

Orchidee

schon klassifiziert

Unbekannt ?

trainiert

System

1. sagt vorher

Output

Es ist eine Orchidee.

2. überprüft

Ergebnis

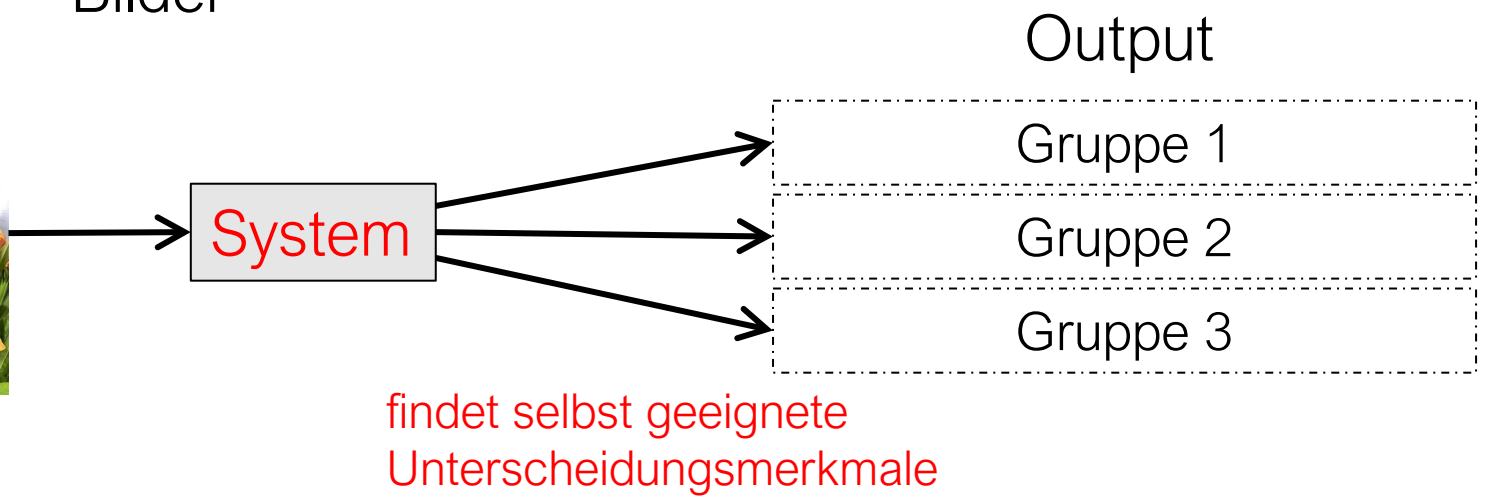
Unsupervised Deep Learning

„Unüberwachtes Lernen“

Beispiel: Algorithmen trainieren Orchideen von Knabenkraut zu unterscheiden.

größeres Datenset→ Input = Bilder

(nicht klassifiziert)



findet selbst geeignete Unterscheidungsmerkmale

Clustering

Clusterverfahren zur Erkennung von Ähnlichkeiten in großen Datenbeständen.

→ Datenpunkte werden Gruppen = „Clustern“ zugeordnet

→ innerhalb einer Gruppe möglichst homogen

→ Komplexität verringern

Unterschied zur Klassifikation beim Supervised Learning:

Daten werden bereits bestehenden Klassen zugeordnet.

Unsupervised Learning
Neue Gruppen identifizieren

Es gibt viele Algorithmen im Clustering.

Supervised Machine Learning

„Überwachtes Lernen“

Beispiel: Algorithmen trainieren Orchideen von Knabenkraut zu unterscheiden.

Datenset→ Input = Bilder

Knabenkraut

Orchidee

schon klassifiziert

trainiert

System

sagt vorher

Output

Es ist eine Orchidee.

Unbekannt ?

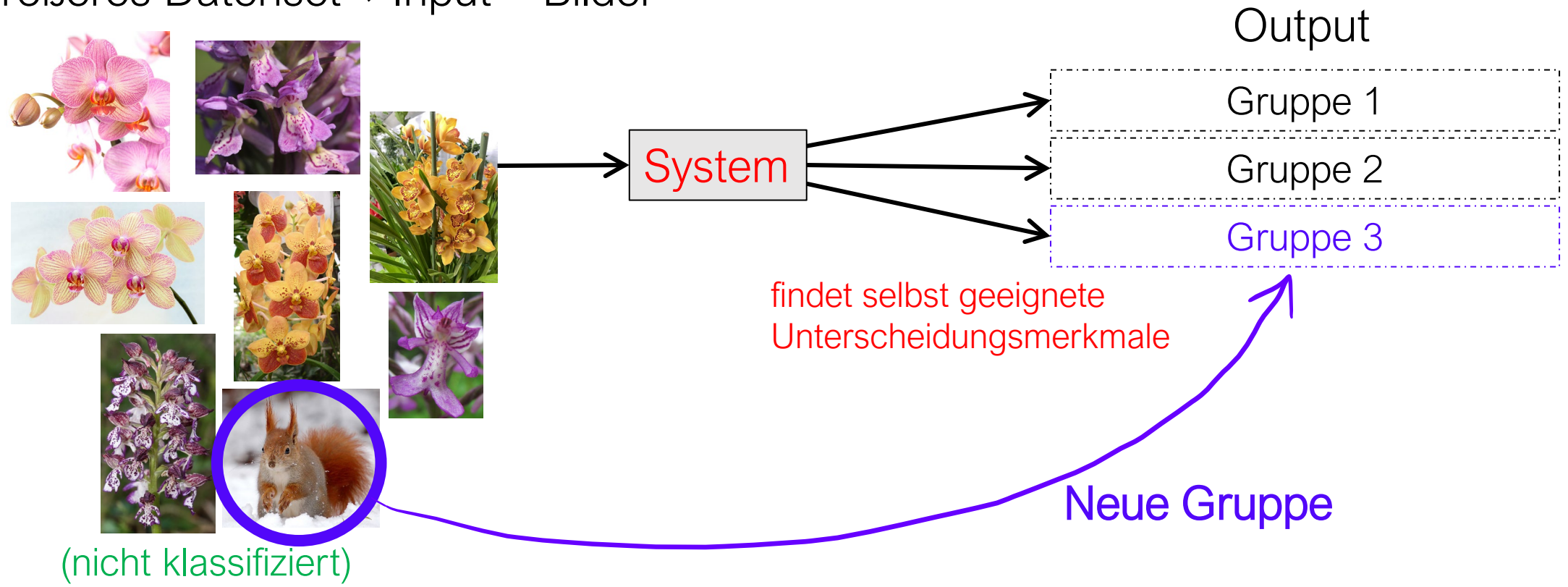
Klassifikation

Unsupervised Deep Learning

„Unüberwachtes Lernen“

Beispiel: Algorithmen trainieren Orchideen von Knabenkraut zu unterscheiden.

größeres Datenset→ Input = Bilder



Mathematische Modellierung

Menge O von Objekten→ repräsentiert als Punkte im Vektorraum $\mathbb{R}^n$

O anhand Indizes ansprechen

n Objekte werden beim Clustern zu einer Partition C_i der Index-Menge $\{1,\ldots,n\}$

$$\rightarrow C_1 \cup \ldots \cup C_i = \{1,\ldots,n\}$$

$$C_l \cap C_j = \emptyset \; für \; l \neq j.$$

Bedeutung: Jeder Datenpunkt wird genau einer Gruppe zugewiesen.

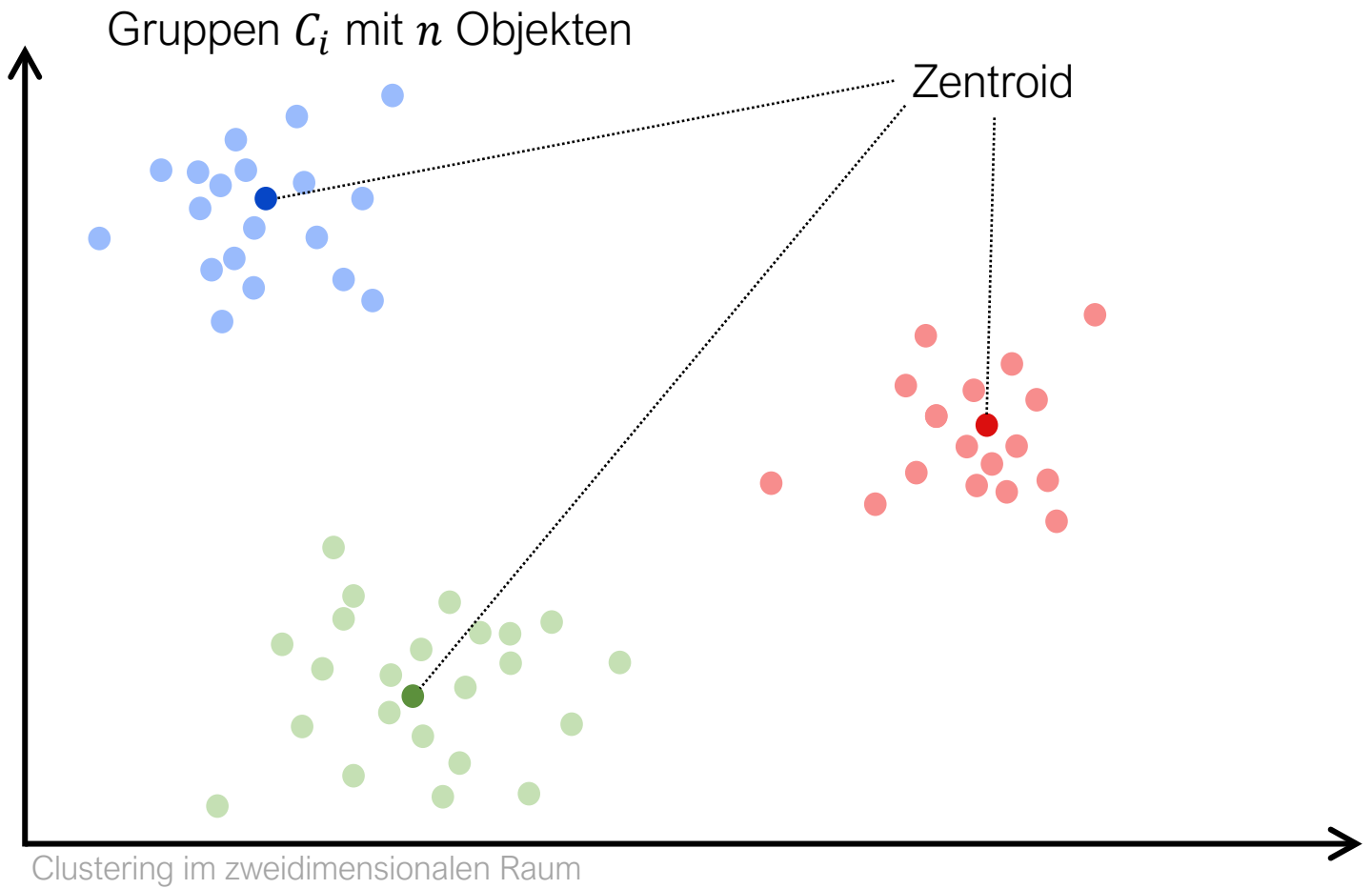
Beispiel: 7 Objekte, 3 Cluster

Dann sind die Teilmengen beispielsweise $\{1,3,6\}, \{2,4\}, \{5,7\}$.

$$\rightarrow \{1,3,6\} \cup \{2,4\} \cup \{5,7\} = \{1,2,3,4,5,6,7\}$$

Beispiel Clustering

hier im $\mathbb{R}^2$



→ disjunkte Aufteilung in 3 Cluster

→ unterschiedliche Aufteilungsverfahren

Disjunktes / Partitionierendes Clustering

k-means Algorithmus

Parameter $k \in \mathbb{N}$ bestimmt die Anzahl der Cluster.

Jedes Cluster C_i wird vom Zentroid $c_i \in \mathbb{R}^n$ repräsentiert.

❶ wähle zufällig *k* Punkte $c_1, \ldots, c_i \in \mathbb{R}^n$ als Zentroide

❷ ordne jeweils alle Objekte $o \in O$ dem nähesten c_i Zentroid zu

❸ Sei C_i das Cluster, d.h. die Menge der Objekte, die c_i zugeordnet wurden.
Berechne ausgehend von C_i den Zentroid c_i neu.

❹ Falls sich in ❸ mindestens ein Zentroid geändert hat, wiederhole ab Schritt ❷,
andernfalls stoppe ⇒ $C_1, \ldots, C_i$ ist eine Partitionierung von O

Ähnlichkeitsmaß

❷ ordne jeweils alle Objekte $\mathbf{o} \in O$ dem nähesten c_i Zentroid zu

→ In einem Cluster C_i liegen Objekte $\mathbf{o}$, die sich ähneln.

Wie wird diese Ähnlichkeit gemessen ?

<u>über eine Distanzfunktion $d(o, c_i)$:</u>

Euklidische Distanz $$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Manhattan Distanz $$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

→ Abstand $\mathbf{0}$ bedeutet beide Objekte sind äquivalent.

Euklidische Distanz

Jeder Datenpunkt ist ein Vektor $x = (x_1, x_2, \ldots, x_n)$, wobei n die Dimension ist.

Beispiel: $n = 2$

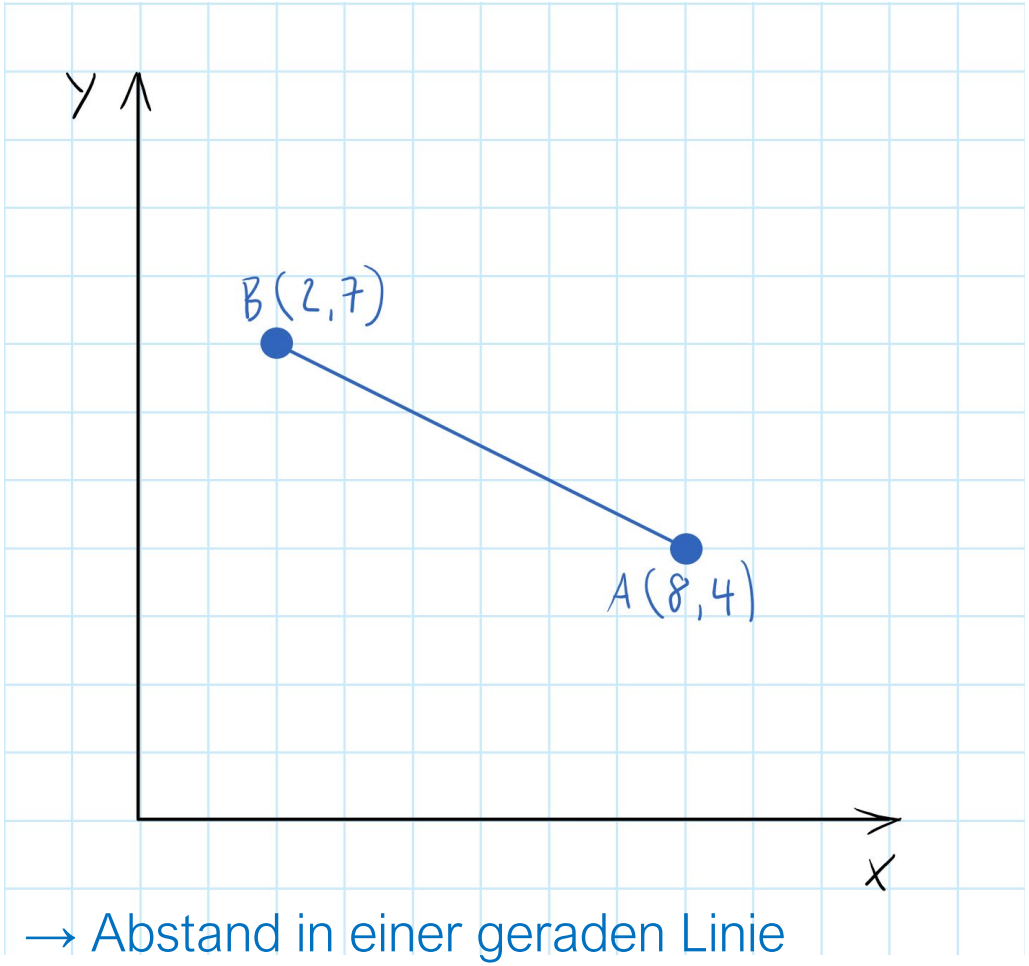
$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$d(A, B) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

$$= \sqrt{(2 - y_1)^2 + (7 - y_2)^2}$$

$$= \sqrt{(2 - 8)^2 + (7 - 4)^2}$$

$$\approx 6{,}71 \text{ LE}$$



→ Abstand in einer geraden Linie

Euklidische Distanz im zweidimensionalen Raum

Grafik angepasst nach: https://www.youtube.com/watch?v=agbcUmOBuyk

Euklidische Distanz

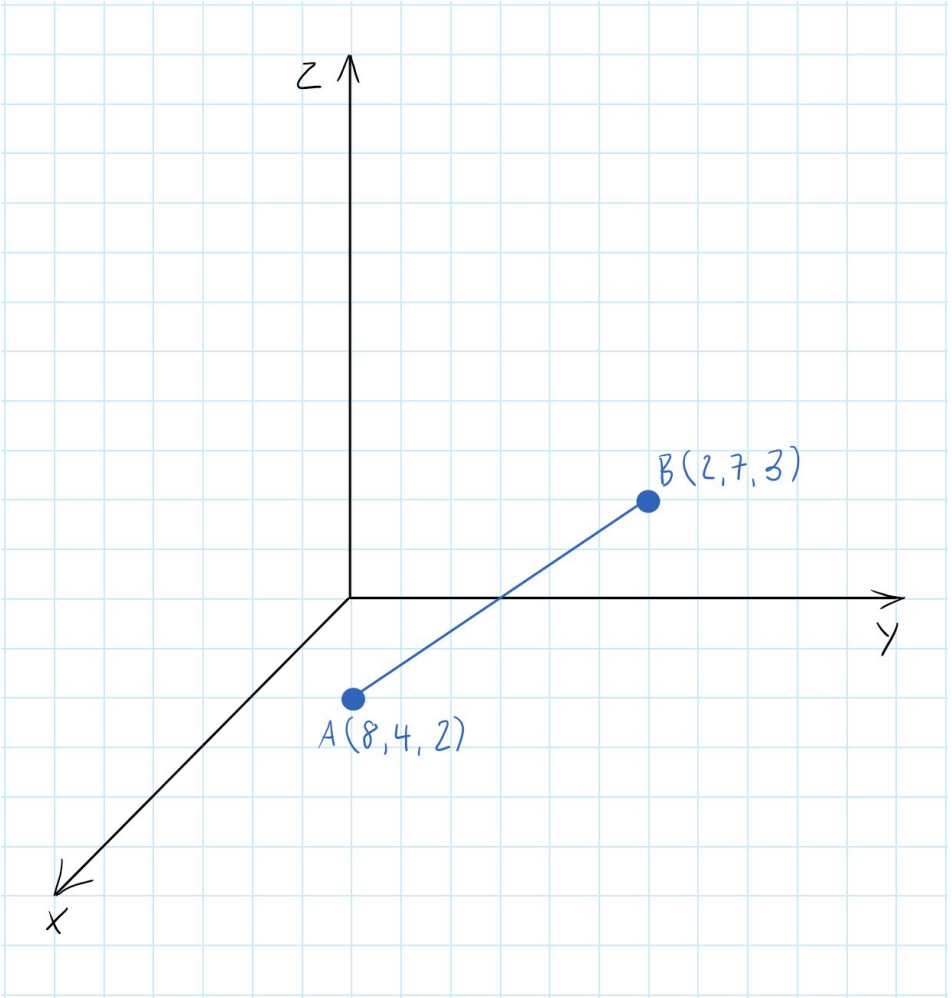
Beispiel: $n = 3$

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$d(A, B) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^3}$$

$$= \sqrt{(8 - 2)^2 + (4 - 7)^2 + (2 - 3)^3}$$

$$\approx 6{,}78 \text{ LE}$$



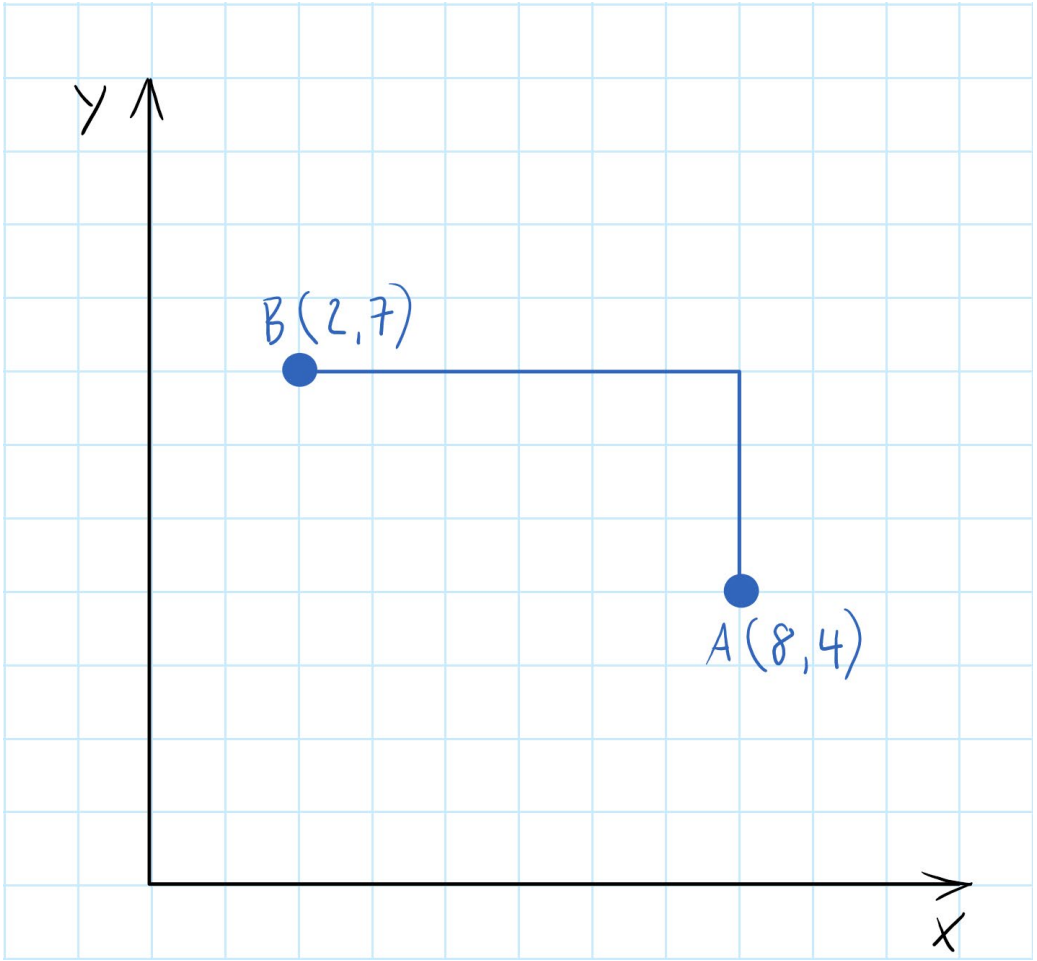
Euklidische Distanz im dreidimensionalen Raum

Manhattan Distanz

Beispiel: $n = 2$

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

$$\begin{aligned} d(A, B) &= |x_1 - y_1| + |x_2 - y_2| \\ &= |2 - 8| + |7 - 4| \\ &= 6 + 3 \\ &= 9 \text{ LE} \end{aligned}$$



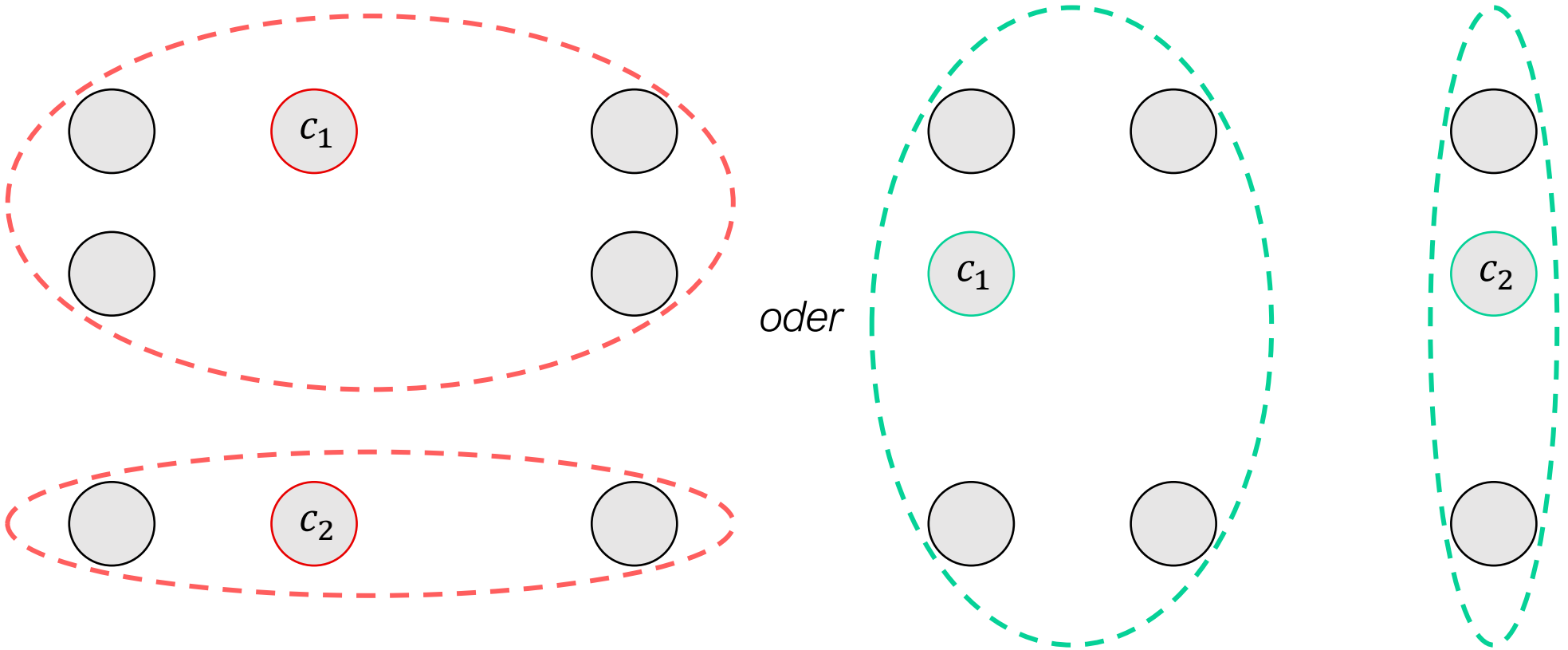
→ Abstand in einer gitterartigen Linie

Manhattan Distanz im zweidimensionalen Raum

Bemerkungen zum *k*-means Algorithmus

Resultat hängt stark von der Zuweisung der anfänglichen Zentroide ab.

→ Abhilfe: Algorithmus mehrfach mit verschiedenen Startzentroiden starten.



Resultate bei verschiedenen Startzentroiden

Grafik angepasst nach: Einführung in das Data Mining Clustering / Clusteranalyse (hpi.de)

Wahl des Parameter *k*

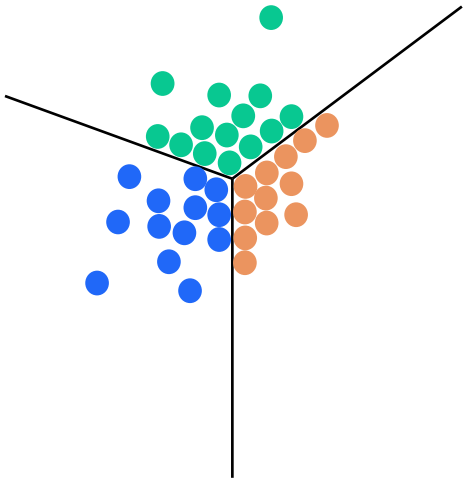
Durch ein falsch gewähltes *k*, resultiert ein schlechtes Clustering.

→ Abhilfe: Algorithmus mehrfach mit verschiedenen Werten für *k* starten.

k sollte nicht zu groß gewählt werden, sonst Overfitting (Überanpassung)

Overfitting: Der Algorithmus erkennt nicht vorhandene Muster im Datenset.

→ falsches Ergebnis



schlechtes Clustering

1 Cluster in 3 Teile aufgeteilt.

→ *k*-means hat nicht die richtige Struktur gefunden

Grafik angepasst nach: Wie man die Nachteile von K-means versteht (qastack.com.de)

Hierarchisches Clustering

top-down Algorithmus

$C = O$, d.h. ein Cluster C in dem alle Objekte $\mathbf{o} \in O$ liegen.

❶ Spalte C in zwei Cluster unter Verwendung des *k*-means Algorithmus mit $k = 2$.

❷ Wähle aus den vorhandenen Clustern C_i ein Cluster C zum Aufspalten aus.

❸ Wiederhole Schritte ❶ und ❷ solange bis
die gewünschte Anzahl an Clustern erreicht ist
oder
jedes Cluster C_i nur ein Objekt $\mathbf{o}$ umfasst.

Hierarchisches Clustering

bottom-up Algorithmus

Wir haben n verschiedene Objekte.

Jedes Objekt $o \in O$ repräsentiert ein Cluster C_i.

❶ Wir starten mit n Clustern, setze $i = n \rightarrow C_n$

❷ bestimmen paarweise alle Abstände zwischen den Clustern

Bestimme die beiden Cluster, die am nähesten zueinander liegen.

vereinige diese zu einem neuen Cluster C

$\rightarrow$ die Clusteranzahl wird um 1 verringert, setze $n = n - 1$

❸ Wenn die gewünschte Anzahl an Clustern erreicht ist

oder $n = 1$, d.h. nur noch ein Cluster existiert, stoppe,

andernfalls gehe zu ❷

Ähnlichkeitsmaß zwischen Clustern

❷ Bestimme die beiden Cluster, die am nähesten zueinander liegen.

vereinige diese zu einem neuen Cluster C

Sei C aus den Clustern C_1 und C_2 entstanden.

Wie wird der Abstand zwischen zwei Clustern berechnet ?

<u>Fusionierungsalgorithmen:</u>

– Single-Linkage

– Complete-Linkage

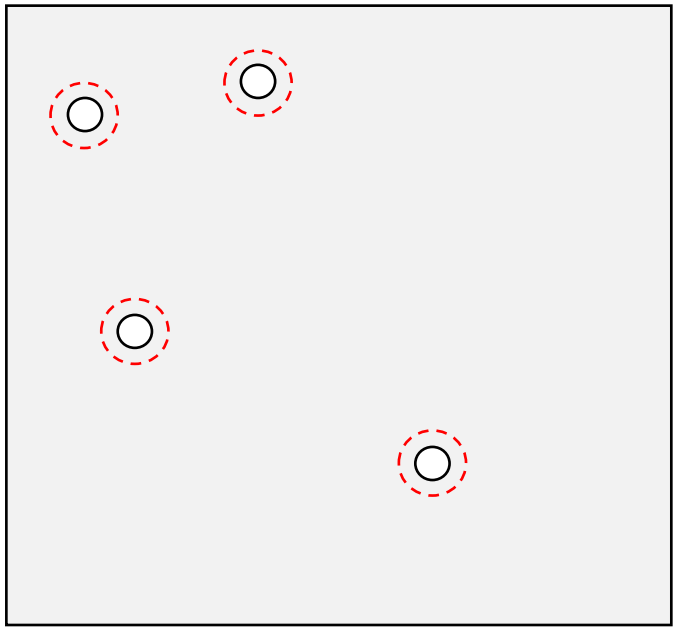
– Average-Linkage

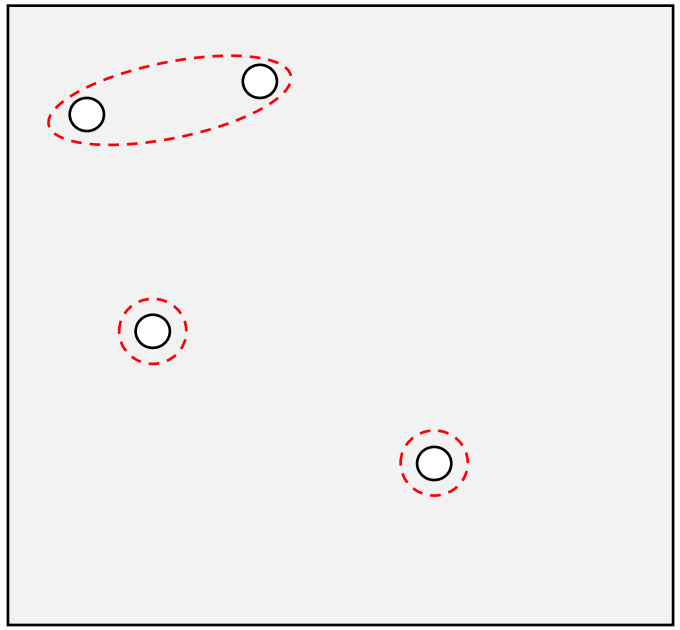
– Centroid-Linkage

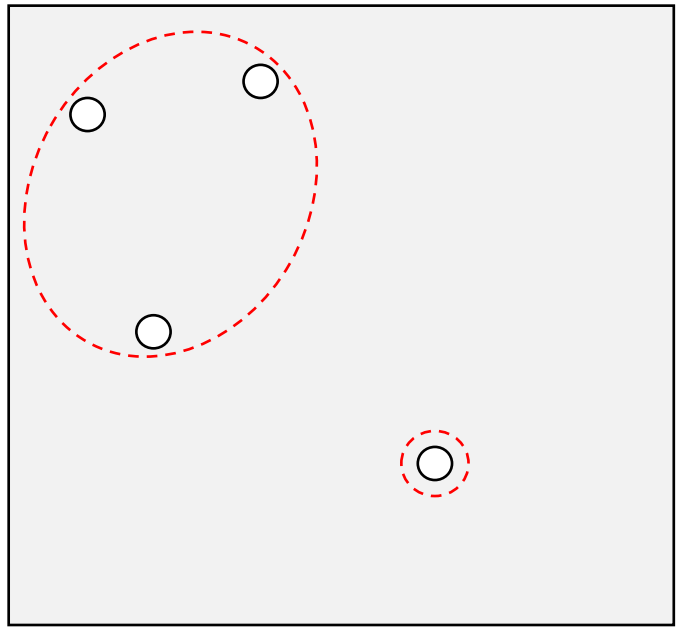
Single-Linkage

❷ bestimmen paarweise alle Abstände zwischen den Clustern

Der kleinste Abstand zwischen zwei Objekten aus verschiedenen Clustern wird als Abstand der Cluster gewählt.



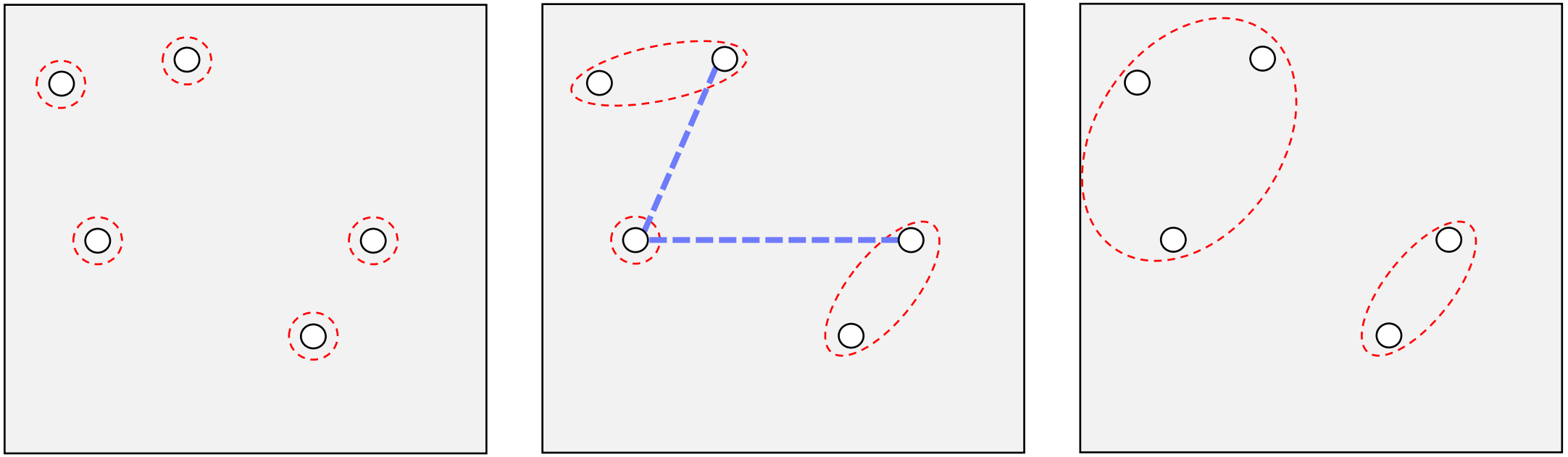




Complete-Linkage

❷ bestimmen paarweise alle Abstände zwischen den Clustern

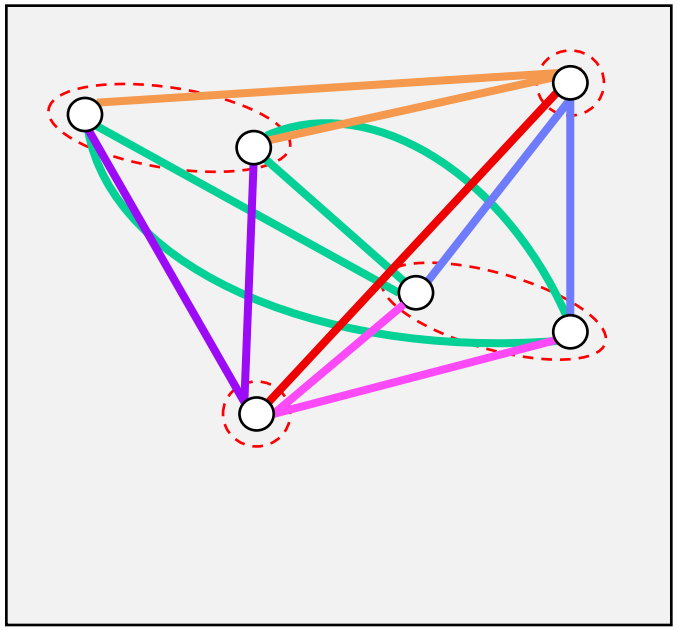
Der größte Abstand zwischen zwei Objekten aus verschiedenen Clustern wird als Abstand der Cluster gewählt. → davon den geringsten Abstand wählen

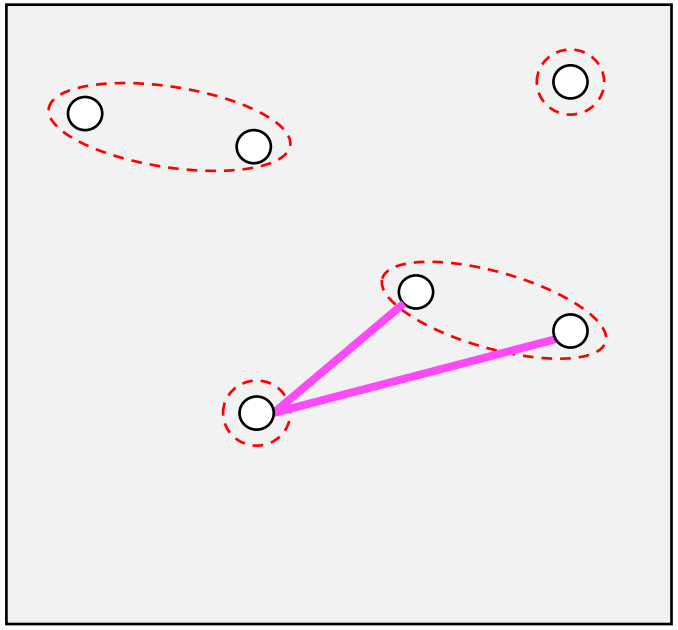


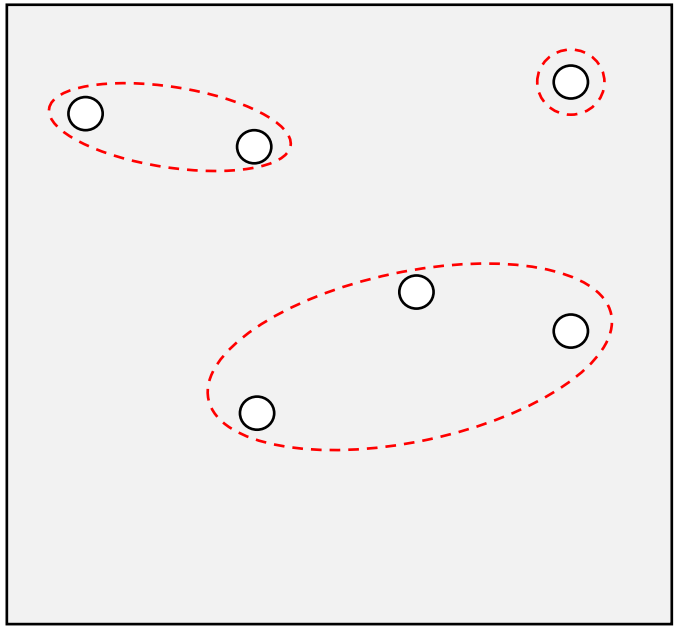
Average-Linkage

❷ bestimmen paarweise alle Abstände zwischen den Clustern

Der mittlere Abstand zwischen Objekten aus zwei verschiedenen Clustern wird als Abstand der Cluster gewählt.







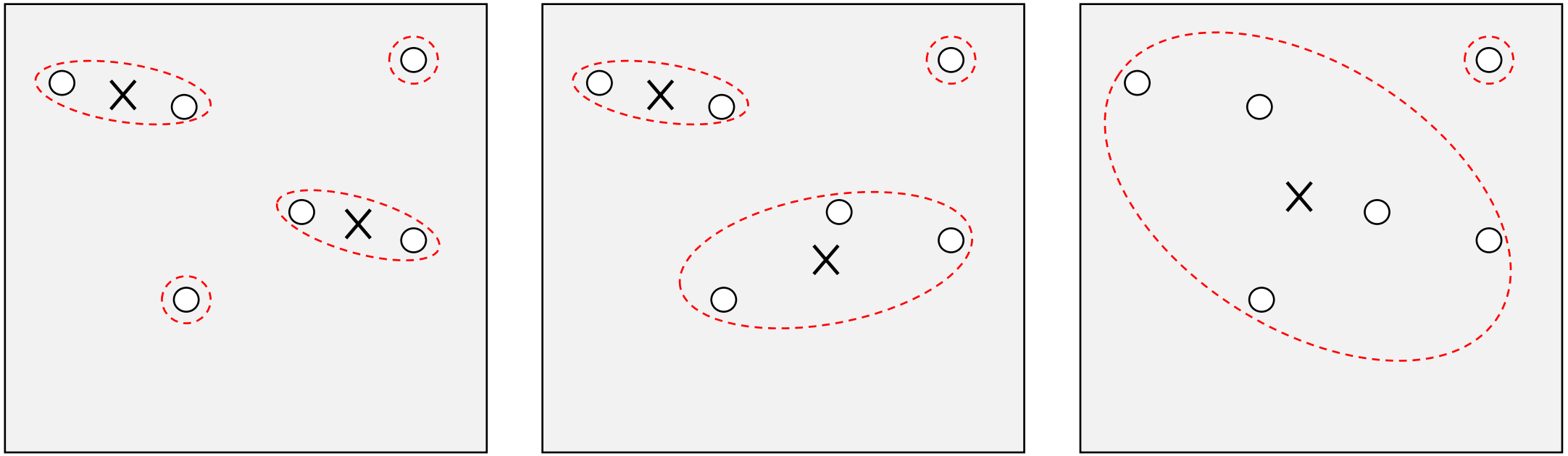
Fusionierungsalgorithmus Average-Linkage

Grafik angepasst nach: Hierarchisches Clustern (tum.de)

Centroid-Linkage

❷ bestimmen paarweise alle Abstände zwischen den Clustern

Der Abstand zwischen den Zentroiden der beiden Cluster wird als Abstand der Cluster gewählt.



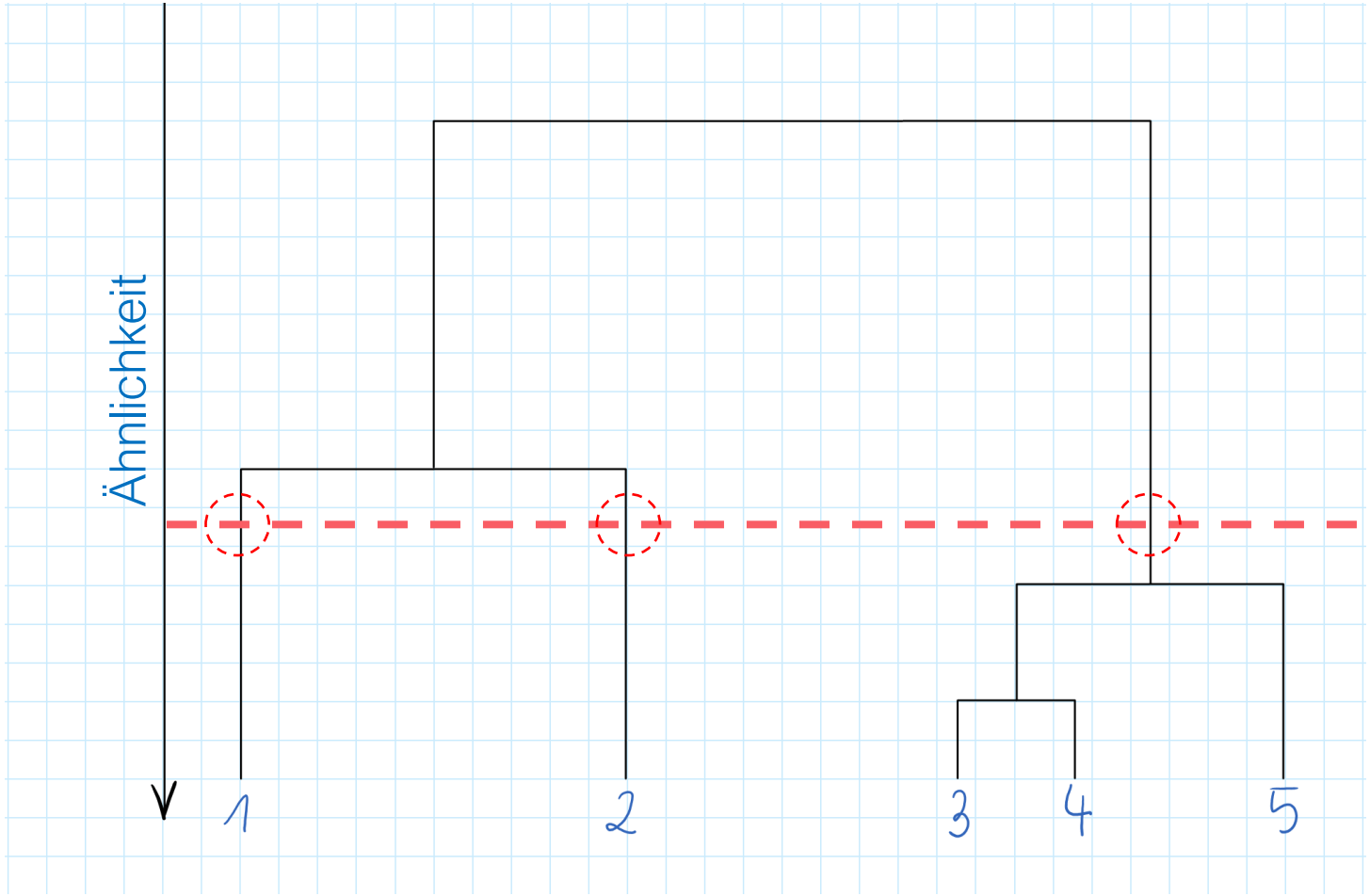
Fusionierungsalgorithmus Centroid-Linkage

Grafik angepasst nach: Hierarchisches Clustern (tum.de)

Dendrogramm

Hierarchisches Clustering liefert Dendrogramm (Visualisierung).

Beispiel:



Visualisierung mittels Dendrogramm

Quelle: https://www.youtube.com/watch?v=agbcUmOBuyk

Einsatzbereiche

– Kundengruppierung

– persönliche Empfehlungen von z.B. Netflix

– Spamfilter

– Betrugserkennung bei Banken

– Bilderkennung und -verarbeitung

– Röntgenbildanalyse

– medizinische Diagnostik

Zusammenfassung

Supervised Machine Learning trifft anhand kategorisierter Daten eine Vorhersage. Soll möglichst mit dem Ziel übereinstimmen.

Unsupervised Deep Learning erkennt eigenständig Merkmale in nicht kategorisierten Daten.

Es gibt viele Algorithmen und alle liefern recht unterschiedliche Resultate.

Distanzfunktionen messen die Ähnlichkeit zwischen zwei Datenpunkten.

Für das Ähnlichkeitsmaß zwischen zwei Clustern existieren Fusionierungsalgorithmen.

Quellen

https://datasolut.com/was-ist-machine-learning/#ueberwachtes-lernen

https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/folien/SS08/Liga/szott_clustering.pdf

https://datasolut.com/wiki/unsupervised-learning/

https://www.youtube.com/watch?v=Q0-DZ_1Y_g4

https://www-m9.ma.tum.de/material/felix-klein/clustering/Methoden/Hierarchisches_Clustern_Beispiel.php

https://qastack.com.de/stats/133656/how-to-understand-the-drawbacks-of-k-means

https://elib.dlr.de/22269/1/beer_diplomarbeit1995.pdf

https://datadrivencompany.de/was-ist-clustering-definition-methoden-und-beispiele/#Definition_von_Clustering_Was_ist_das

https://de.wikipedia.org/wiki/Clusteranalyse#Partitionierende_Clusterverfahren

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

https://de.wikipedia.org/wiki/Hierarchische_Clusteranalyse#Distanz-_und_%C3%84hnlichkeitsma%C3%9Fe

https://www-m9.ma.tum.de/material/felix-klein/clustering/Allgemeines/Abstandsmasse.php

https://studyflix.de/mathematik/euklidische-distanz-1972

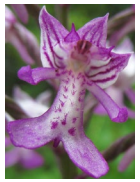
https://www.ifad.de/hierarchische-clusteranalyse/

Bildquellen

Quelle: 12621226-Phalaenopsis-Salmion.jpg (1100×1100) (stockfood.com)

Quelle: a258c645ae999e3a8e545ea9cbb58e52.jpg (1500×1001) (pinimg.com)

Quelle: 35615581610_768ab74270_o.jpg (1600×1050) (staticflickr.com)



Quelle: R5acb4528240ab7ae8443dc4c897686f6 (1244×933) (bing.com)

Quelle: Human Icon - Bing

Quelle: eichhoernchen_im_winter.jpg (1380×920) (schoepfung.eu)

Quelle: B33877A1-06DD-4CBA-B627-98BB524854EA-scaled.jpeg (1920×2560) (vivaioromagarden.

Quelle: 3080157151_af230fcfb7_b.jpg (768×1024) (staticflickr.com)

Quelle: 1920px-Lady_orchid_Orchis_purpurea_inflorescence.jpg (1920×2891) (wikimedia.org)