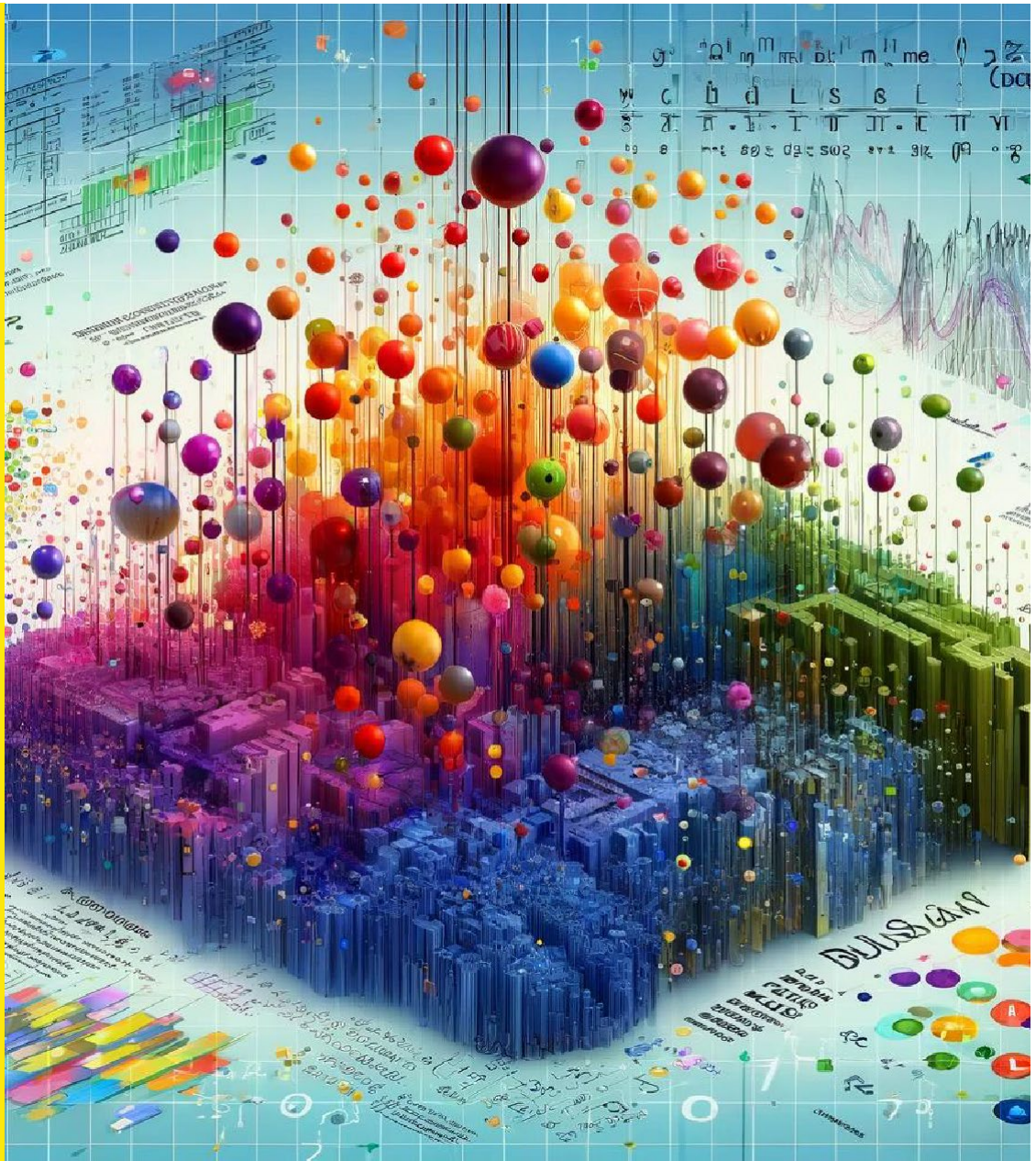
University of Applied Sciences and Arts Northwestern Switzerland
School of Business

swissuniversities

AACSB ACCREDITED

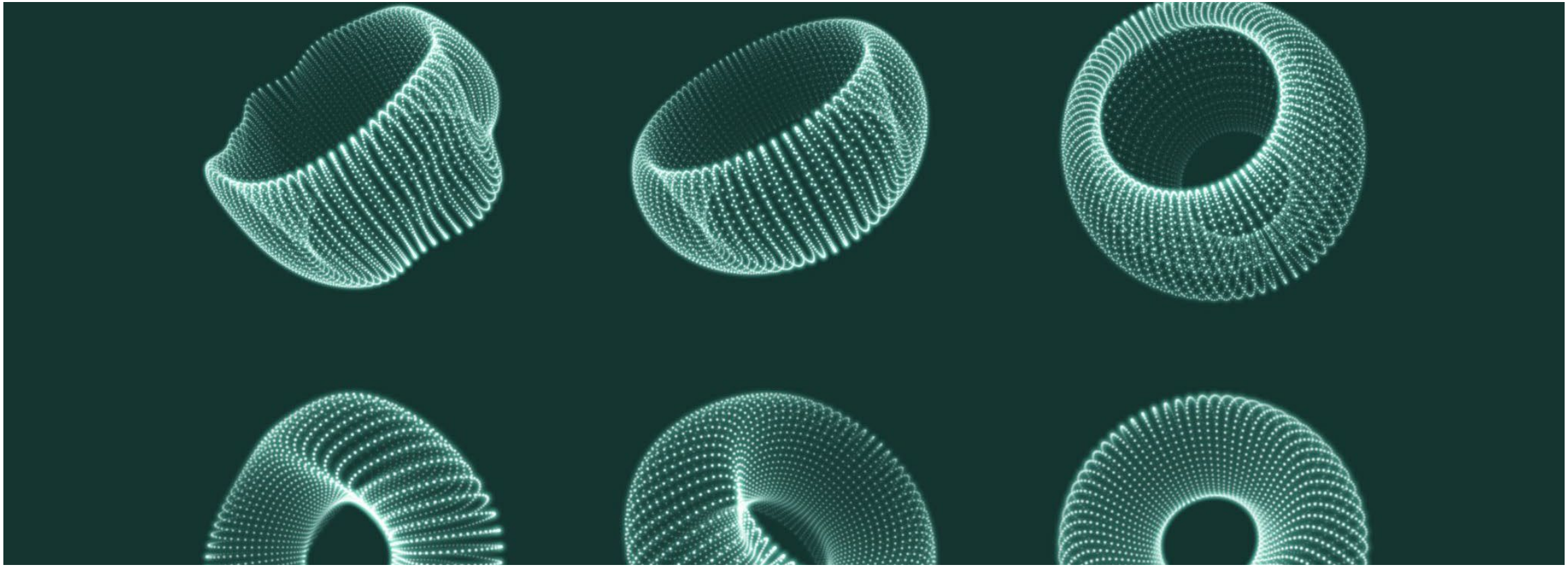
Unsupervised Learning

Maschinelles Lernen | BSc BAI

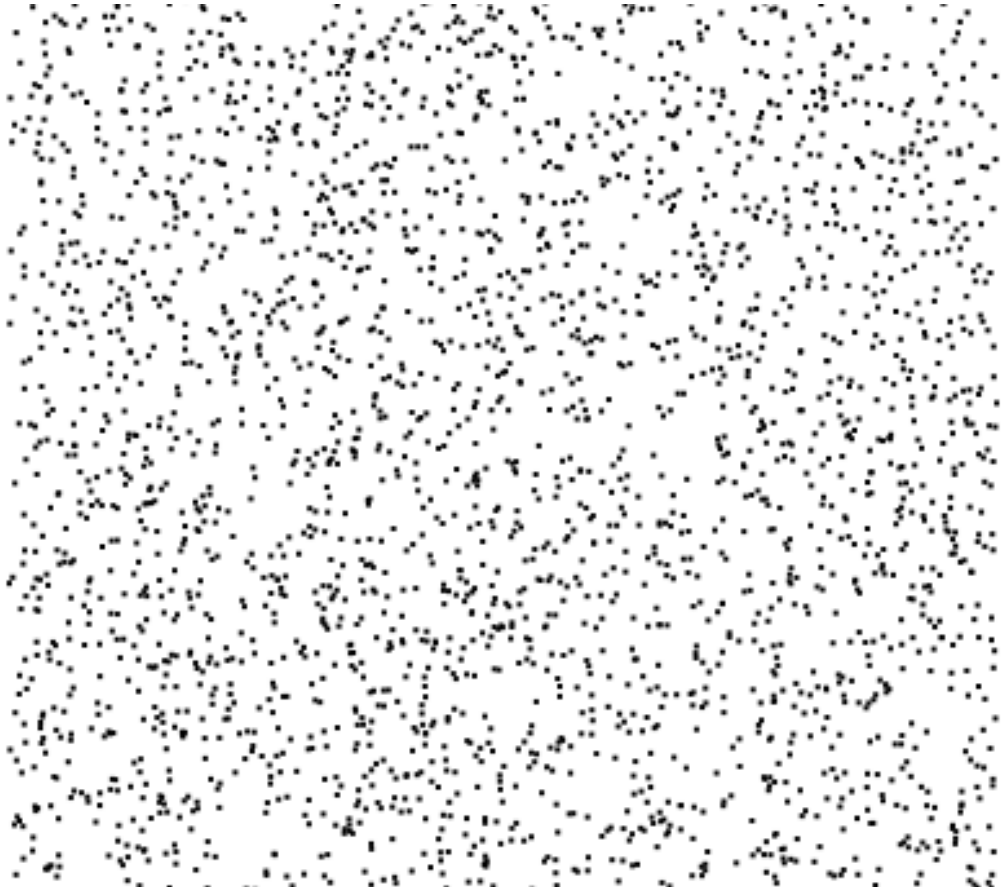


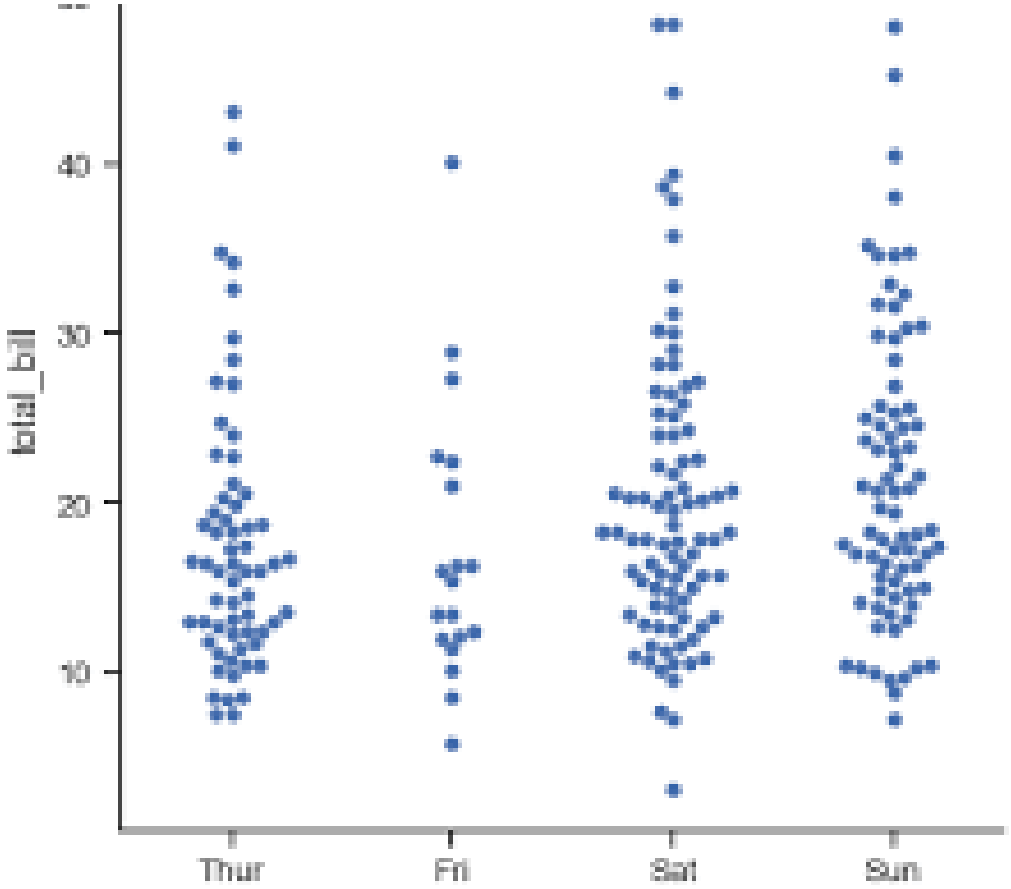
Prof. Dr. Manuel Renold & Prof. Dr. Andreas Martin

Varianten des unüberwachten Lernens I



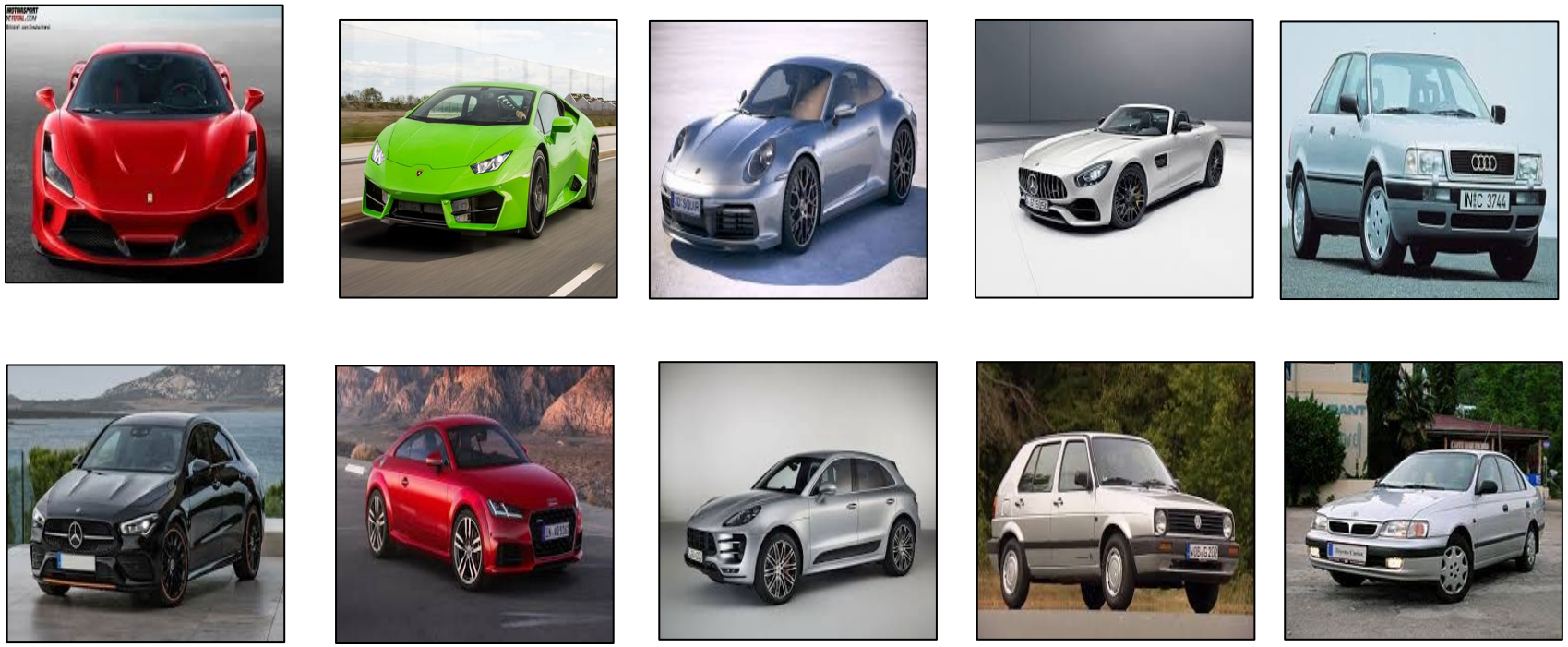
Unsupervised Learning





Clustering

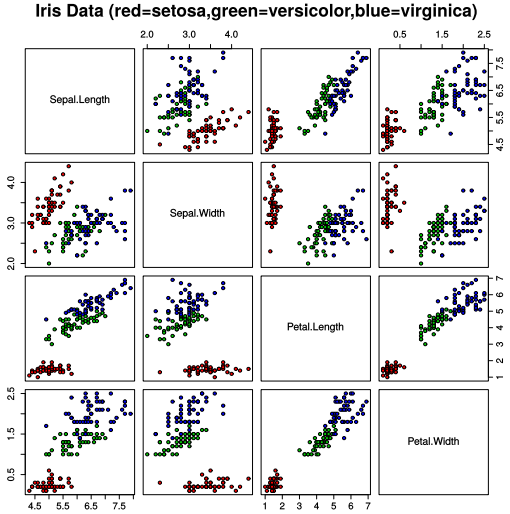
Clustering is the task of ***grouping a set of objects*** in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters)



Dimensionsreduktion

Dimensionality Reduction is the task of transforming the data into an optimal representation of lower dimensionality
- –for visualization (normally 2D, at most 3D)
- –to generate better features for other learning algorithms

Example:



4 Variables:
- sepal width & length
- petal width & length

For Example:

Dimensionality reduction by

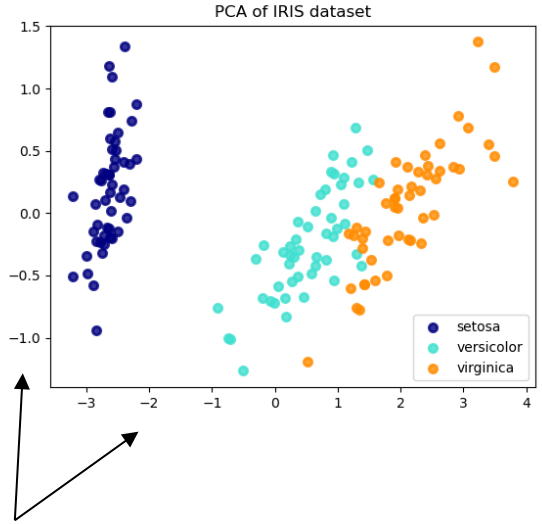
Principal Component Analysis

identifies the directions with
1. largest
2. second largest

…

variance of the data.

→ dimensions which contain most of the information

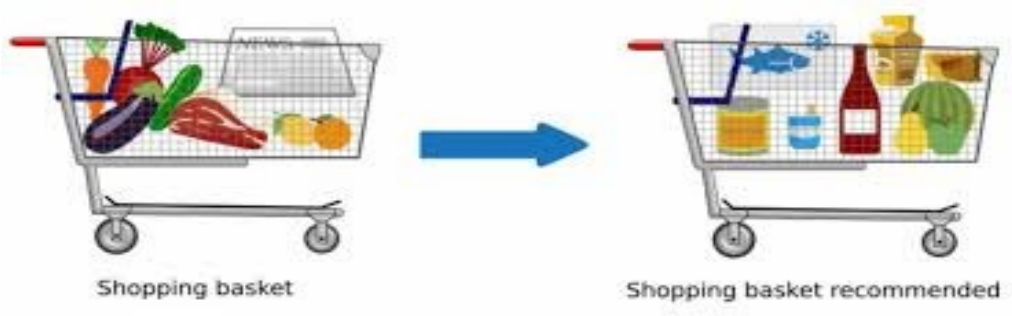


the two principal components

Assoziation Rule Mining

The goal of **Association Rule Mining** is to find associations between the different items that customers place in their shopping baskets, in order to develop marketing strategies according to which items are frequently purchased together

- Where should products be placed in the store to optimize their sales?
- Which products are bought together? change over time?
- Does the brand make a difference?
- How are the demographics of the neighborhood affecting what customers are buying?

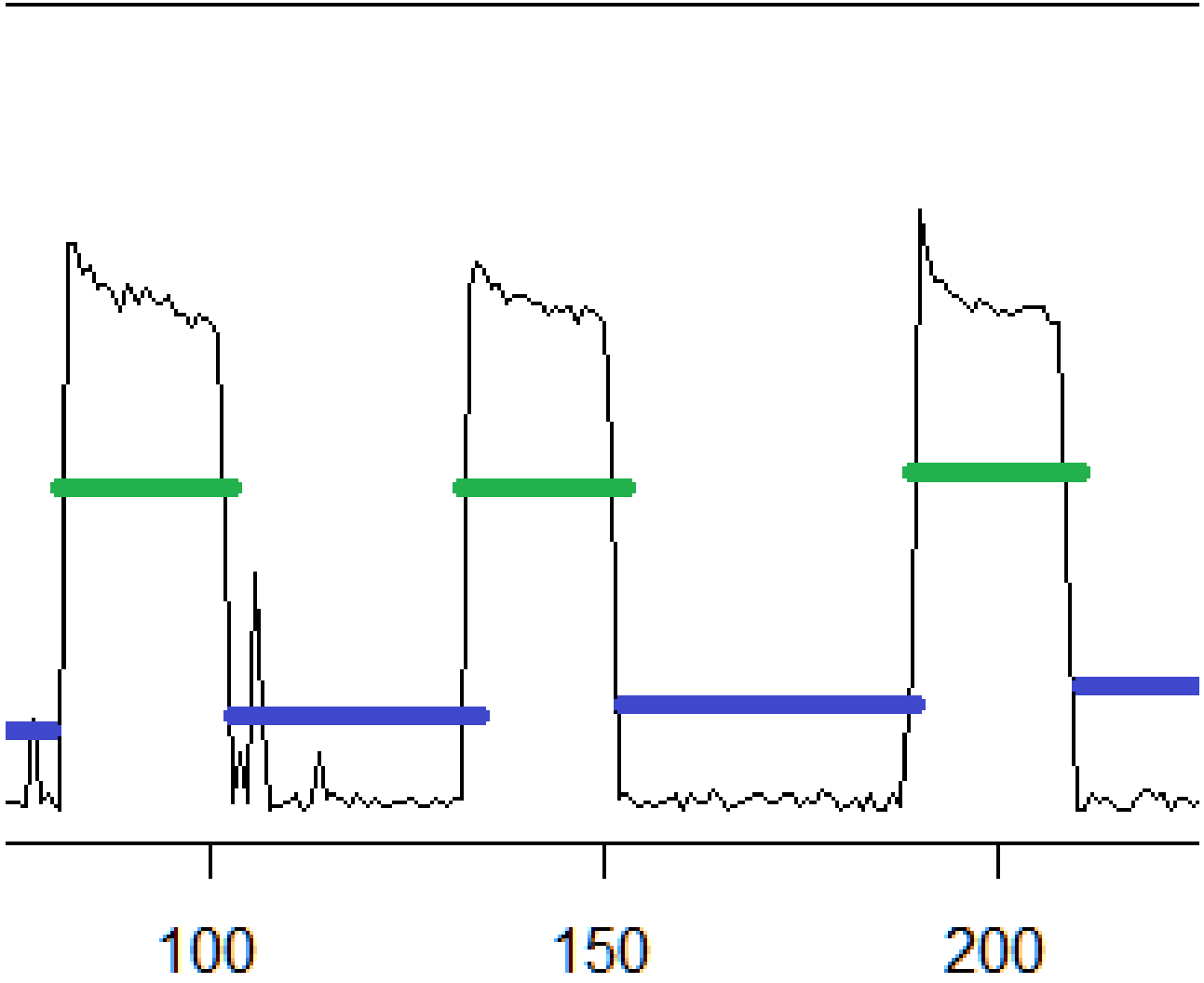


Outlier Detection

The tasks in **Anomaly/Outlier Detection** is to identify samples that are very different from the typical examples in the reference dataset

Possible applications are: Identification of

- Patterns that indicate upcoming machine failures
- Fraudulent financial transactions
- Computer network intrusion
- …



Deep Learning

Applies to both, supervised and unsupervised learning

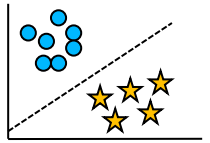
Feature-based Machine Learning

Feature Extraction
(SIFT, SURF, LBP, HOG, etc.)

(0.2, 0.4, …)

(0.4, 0.3, …)

Algorithms: SVM, Decision Tree etc.

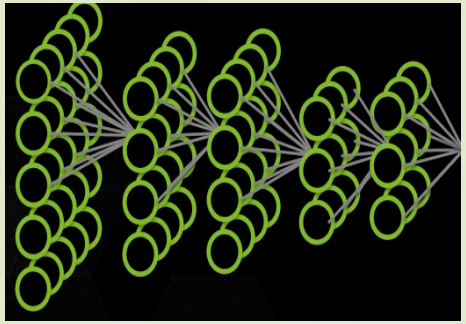


Container Ship

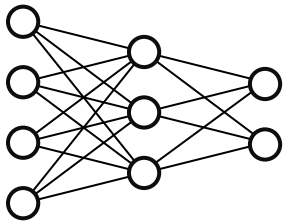
Tiger

Deep Learning

takes raw pixels, features are learned by the network!



Algorithms: Neural Networks



Container Ship

Tiger

Clusteranalyse

Einführung in die Clusteranalyse

Einführung in die Clusteranalyse – mit einem Beispiel

- Beispiel:
  - Motivation: Herausfinden, warum Einladungen zu Aktivitäten unterschiedliche Zusagensraten haben.
  - Methode: Durchführen einer Umfrage zu Präferenzen bezüglich zweier verschiedener Aktivitätsarten.
  - Beobachtung: Entstehung von zwei Gruppen mit unterschiedlichen Interessen.

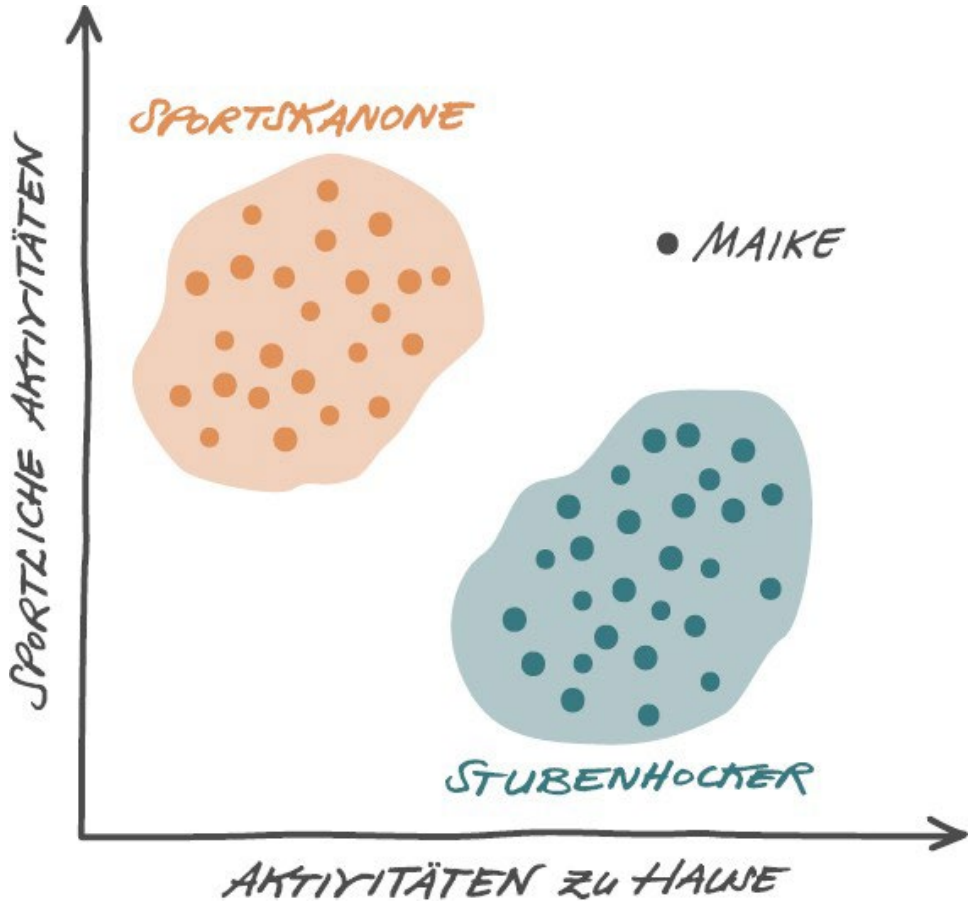
- Definition: Clusteranalyse ist ein unüberwachtes Lernverfahren, das Objekte basierend auf Ähnlichkeiten in Gruppen einteilt, um die Unterschiede zwischen den Gruppen zu maximieren.

NAME	AKTIVITÄTEN ZU HAUSE	SPORTLICHE AKTIVITÄTEN
JANNIK	1	8
NANINA	10	3
MICHAEL	2	9
MAIKE	8	8
ALEXANDROS	9	1
JOHANNES	4	10
...	...	...

Aberham, J., & Kuruc, F. (2019). Clusteranalyse. In K. Kersting, C. Lampert, & C. Rothkopf (Eds.), Wie Maschinen lernen (pp. 53–57). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26763-6_7

Beispiel: Anwendung der Clusteranalyse auf soziale Gruppen

- Problem: Unterschiedliche Präferenzen führen zu nicht einheitlichen Zusagen bei Gruppenaktivitäten.
- Lösungsansatz: Einteilung in Gruppen basierend auf ihren Präferenzen zu Aktivitäten.
- Ergebnis: Erhöhung der Zusagensrate durch angepasste Aktivitätseinladungen entsprechend der Gruppenpräferenzen.
- Herausforderung: Schwierigkeiten bei der eindeutigen Zuordnung einzelner Personen zu den definierten Gruppen.



Aberham, J., & Kuruc, F. (2019). Clusteranalyse. In K. Kersting, C. Lampert, & C. Rothkopf (Eds.), Wie Maschinen lernen (pp. 53–57). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26763-6_7

Ziele und Herausforderungen der Clusteranalyse

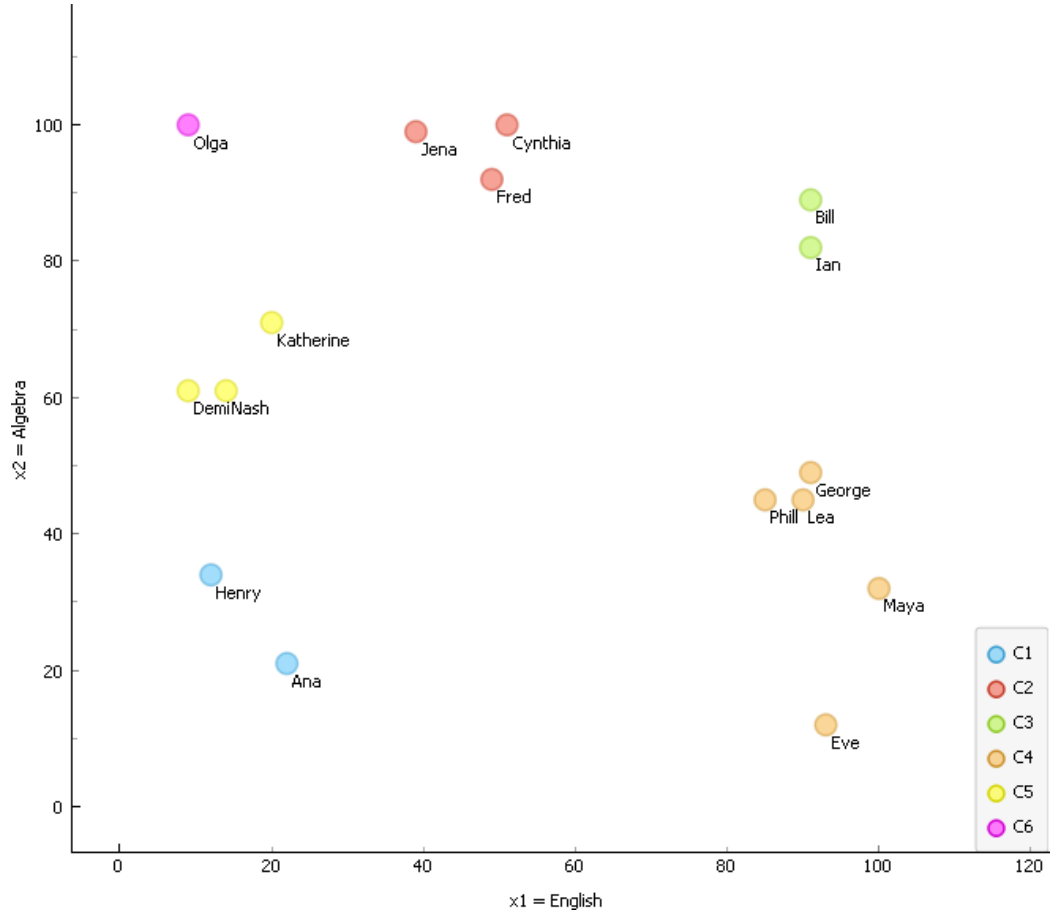
- Zweck: Identifikation von Ähnlichkeitsmustern in Daten zur Gruppenbildung und Mustererkennung.
- Merkmale des unüberwachten Lernens: Keine vorherige Kenntnis über die Anzahl oder Art der Gruppen.
- Anwendungsgebiete: Von Marketingsegmentierungen bis hin zur Analyse von sozialen Netzwerken.
- Herausforderungen: Bestimmung der Anzahl von Clustern, Umgang mit Daten, die nicht klar zugeordnet werden können.
- Nutzen: Ermöglicht tiefergehende Einsichten in Datenstrukturen für gezielte Strategien und Entscheidungen.

Aberham, J., & Kuruc, F. (2019). Clusteranalyse. In K. Kersting, C. Lampert, & C. Rothkopf (Eds.), Wie Maschinen lernen (pp. 53–57). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26763-6_7

Clustering und Euklidische Distanz

Clustering und Euklidische Distanz

- Clustering ermöglicht die Entdeckung natürlicher Gruppierungen in Daten.
  - Kann zur Segmentierung von Nutzerprofilen, Analyse von Kaufverhalten oder Identifizierung ähnlicher medizinischer Muster verwendet werden.
- Euklidische Distanz: Eine direkte Linie zwischen zwei Punkten im mehrdimensionalen Raum.
  - Eignet sich besonders für quantitative und kontinuierliche Daten.
  - Fundamentales Konzept in vielen Clustering-Algorithmen, einschliesslich k-Means und hierarchisches Clustering
  - Vereinfacht die Komplexität grosser Datensätze durch Bildung verständlicher Gruppen.



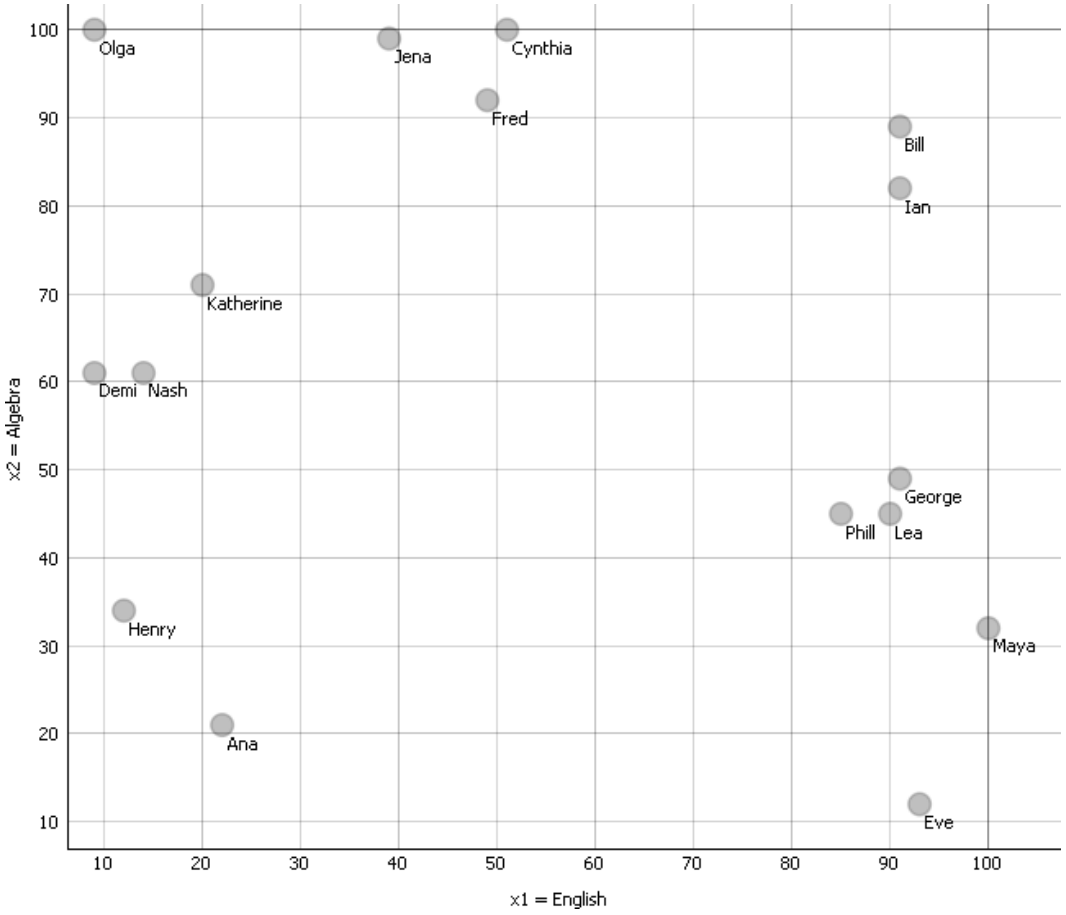
Datenauswahl und Vereinfachung

- Die Auswahl repräsentativer Merkmale ist entscheidend für die Effektivität des Clusterings.
- Beispiel Schülerleistungen: Fokus auf Englisch und Algebra zur Vereinfachung.
- Reduzierung der Dimensionalität erleichtert die Visualisierung und Analyse.
- Wichtigkeit der Datenvorbereitung: Reinigung und Standardisierung der Daten vor dem Clustering.
- Die Selektion der Merkmale beeinflusst direkt die Interpretierbarkeit der Cluster.

	Student	Algebra	English	French	History	Biology	Physics	Physical
1	Ana	21	22	30	32	37	46	99
2	Bill	89	91	95	65	39	11	29
3	Cynthia	100	51	89	21	70	100	27
4	Demi	61	9	15	18	100	90	8
5	Eve	12	93	99	39	47	17	63
6	Fred	92	49	17	17	70	98	73
7	George	49	91	99	97	96	81	69
8	Henry	34	12	30	32	12	33	96
9	Ian	82	91	80	20	93	87	22
10	Jena	99	39	18	19	97	77	23
11	Katherine	71	20	50	10	99	78	12
12	Lea	45	90	100	45	20	15	100
13	Maya	32	100	98	97	72	22	37
14	Nash	61	14	4	15	42	51	39
15	Olga	100	9	22	8	11	92	29
16	Phill	45	85	90	100	38	92	21

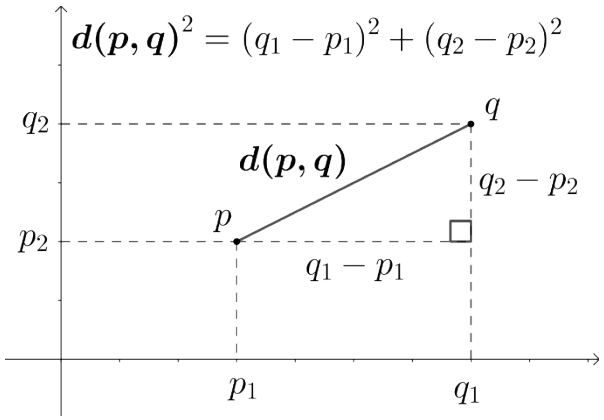
Visualisierung im Streudiagramm

- Streudiagramme (Scatter Plot) ermöglichen eine intuitive Einschätzung der Datenstruktur.
- Label erleichtern die Identifikation von Ausreissern und Mustern.
- Beobachtungen liefern erste Anhaltspunkte für die Bildung von Clustern.
- Die räumliche Nähe im Diagramm deutet auf ähnliche Merkmalsausprägungen hin.
- Visualisierung ist ein mächtiges Werkzeug zur Hypothesenbildung im Data Mining.



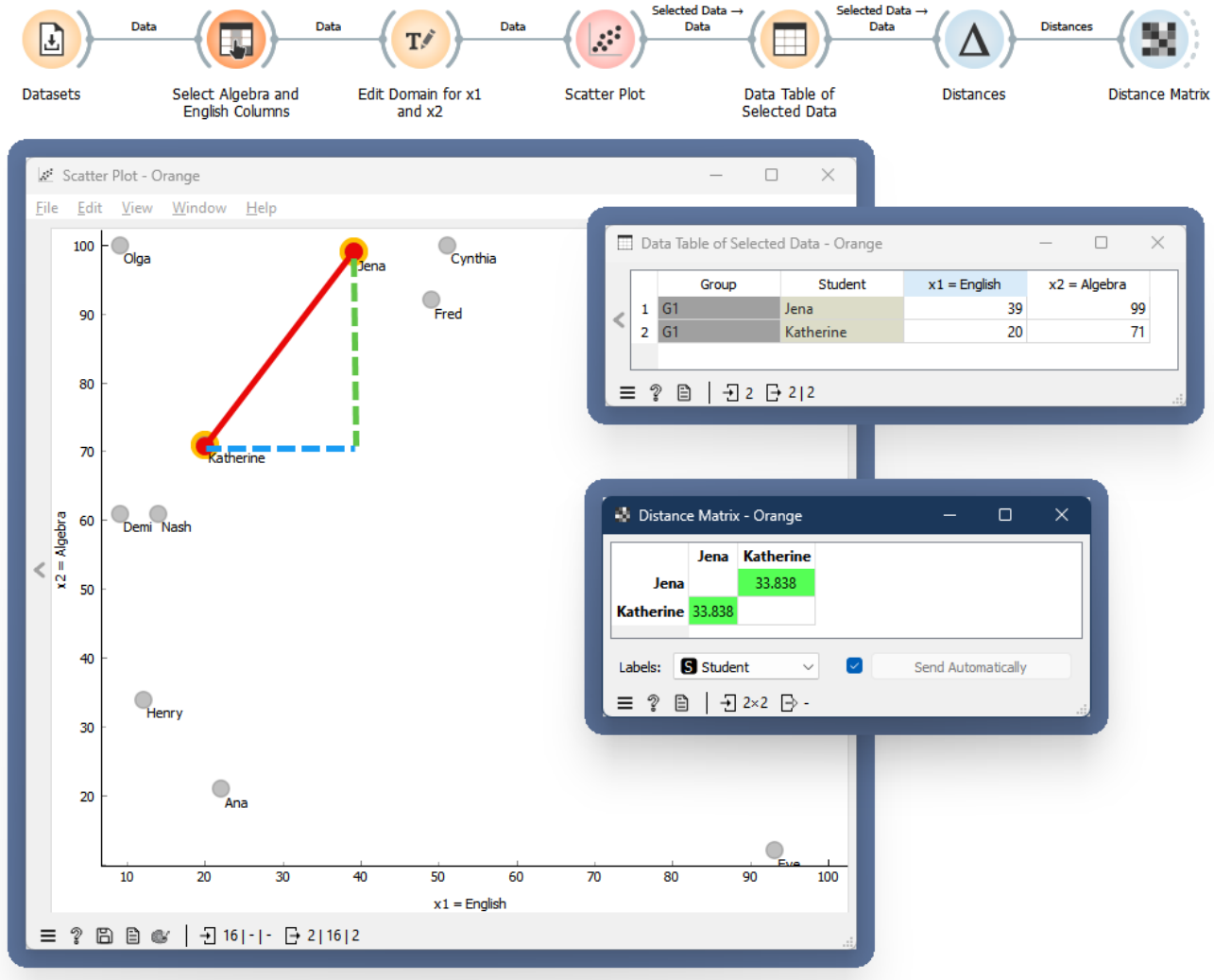
Euklidische Distanz – Messung der Distanz zwischen Datenpunkten

- Die Euklidische Distanz misst den "direkten" Abstand zwischen zwei Punkten, unabhängig von der Dimensionalität des Raumes.
- $Euclidean\ Dinstance\ (d) = \sqrt{(q_{x1} - p_{x1})^2 + (q_{x2} - p_{x2})^2}$
- Grundlage für die Beurteilung der Ähnlichkeit zwischen Objekten.
- Bietet eine objektive Methode zur Quantifizierung von Unterschieden.
- Ermöglicht die Transformation von qualitativen Merkmalsunterschieden in quantifizierbare Distanzen.



Berechnungsbeispiel der Euklidischen Distanz

- Die Berechnung integriert beide Merkmalsunterschiede (Englisch $x1$ und Algebra $x2$) in einem einzigen Distanzwert d.
  - $d = \sqrt{(q_{x1} - p_{x1})^2 + (q_{x2} - p_{x2})^2}$
- Beispiel Distanz von Katherine K und Jena J:
  - $d(K, J) = \sqrt{(J_{x1} - K_{x1})^2 + (J_{x2} - K_{x2})^2}$
- Punkte: Katherine $K\,(20, 71)$ & Jena $J(39, 99)$:
  - $33.838 = \sqrt{(39 - 20)^2 + (99 - 71)^2}$



Anwendung der Euklidischen Distanz im Clustering

- Die Euklidische Distanz bildet das Fundament für die quantitative Analyse von Ähnlichkeiten.
- Erleichtert die Dateninterpretation durch Bildung von intuitiv verständlichen Gruppen.
- Vorbereitung auf weiterführende Clustering-Methoden und -Algorithmen.
- Eröffnet Möglichkeiten für datengetriebene Entscheidungen und Erkenntnisse.
- Anwendbar in diversen Feldern von der Kundenanalyse bis hin zur genetischen Forschung.

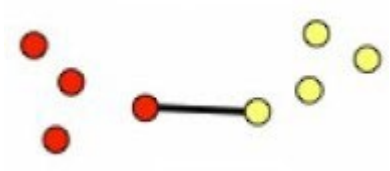
Hierarchisches Clustering

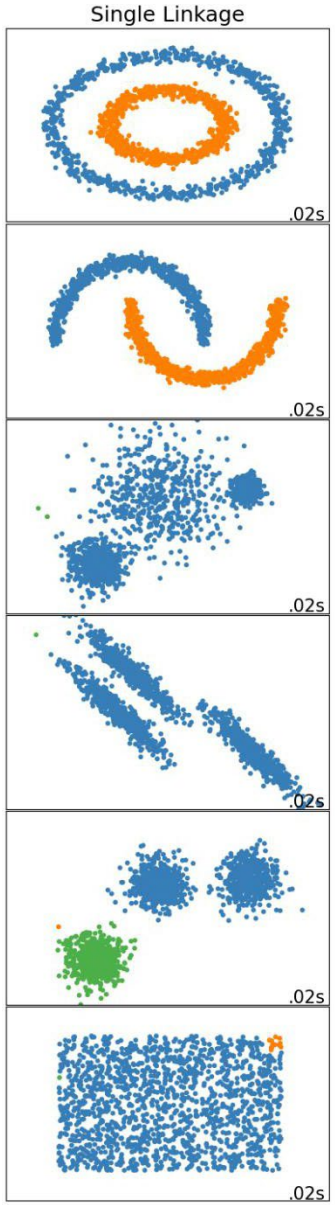
Hierarchisches Clustering

- Hierarchisches Clustering organisiert Datenpunkte in einer hierarchischen Baumstruktur.
- Beginnt mit jedem Datenpunkt als eigenständigem Cluster.
- Führt schrittweise die am nächsten liegende Cluster zusammen, basierend auf einem Distanzmass.
- Ermöglicht die Betrachtung von Daten auf verschiedenen Ebenen der Aggregation.
- Euklidische Distanz oft genutzt zur Messung der Nähe zwischen Punkten.
- Die Wahl des Distanzmasses beeinflusst die Clusterstruktur.

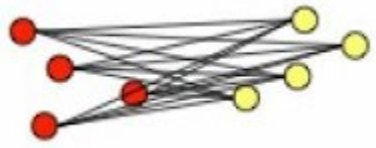
Distanzmasse im Hierarchischen Clustering (1)

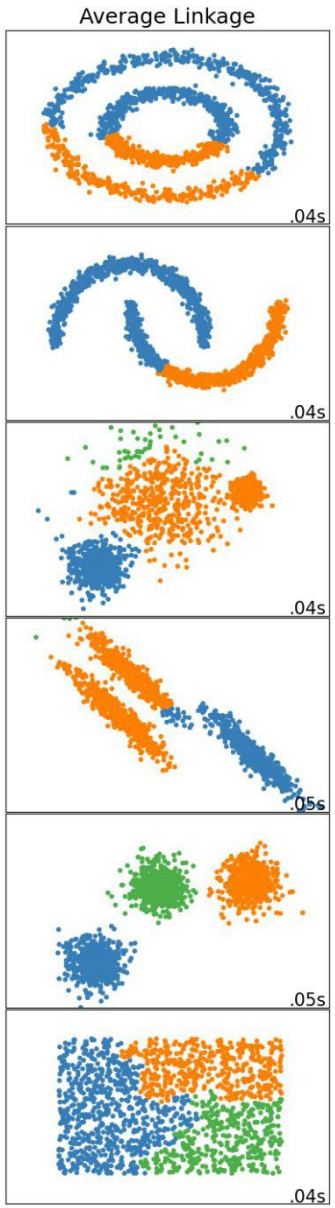
- **Single Linkage** (Nächster Nachbar): Misst die Distanz zwischen den nächsten Mitgliedern zweier Cluster.
  - Betrachtet die kürzeste Distanz zwischen irgendwelchen zwei Punkten in verschiedenen Clustern.
  - Neigt dazu, "Ketten" zu bilden, die sich weit in den Raum erstrecken können.
  - Empfindlich gegenüber Ausreissern, da eine kleine Distanz zwischen zwei Punkten Cluster verbinden kann.
  - Geeignet für die Identifizierung von nicht-sphärischen Datenstrukturen.





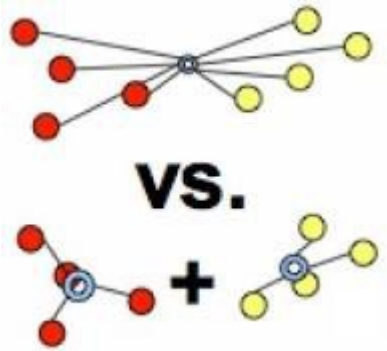
- **Average Linkage** (Durchschnittsbindung): Berechnet die durchschnittliche Distanz zwischen allen Mitgliedern zweier Cluster.
  - Berechnet die durchschnittliche Distanz zwischen allen Paaren von Punkten in zwei Clustern.
  - Fördert die Bildung von Clustern, die intern kohäsiver und extern gut getrennt sind.
  - Weniger anfällig für Ausreisser als Single Linkage.
  - Bietet oft ein ausgewogeneres Dendrogramm.

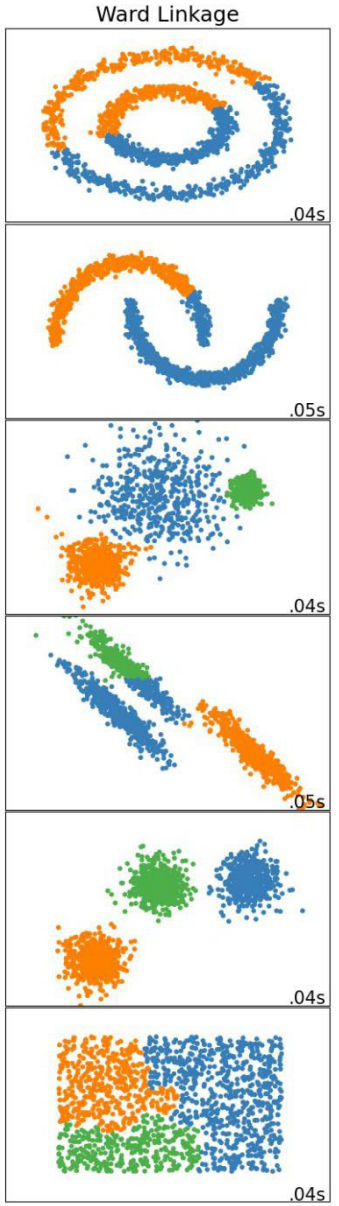




Distanzmasse im Hierarchischen Clustering (2)

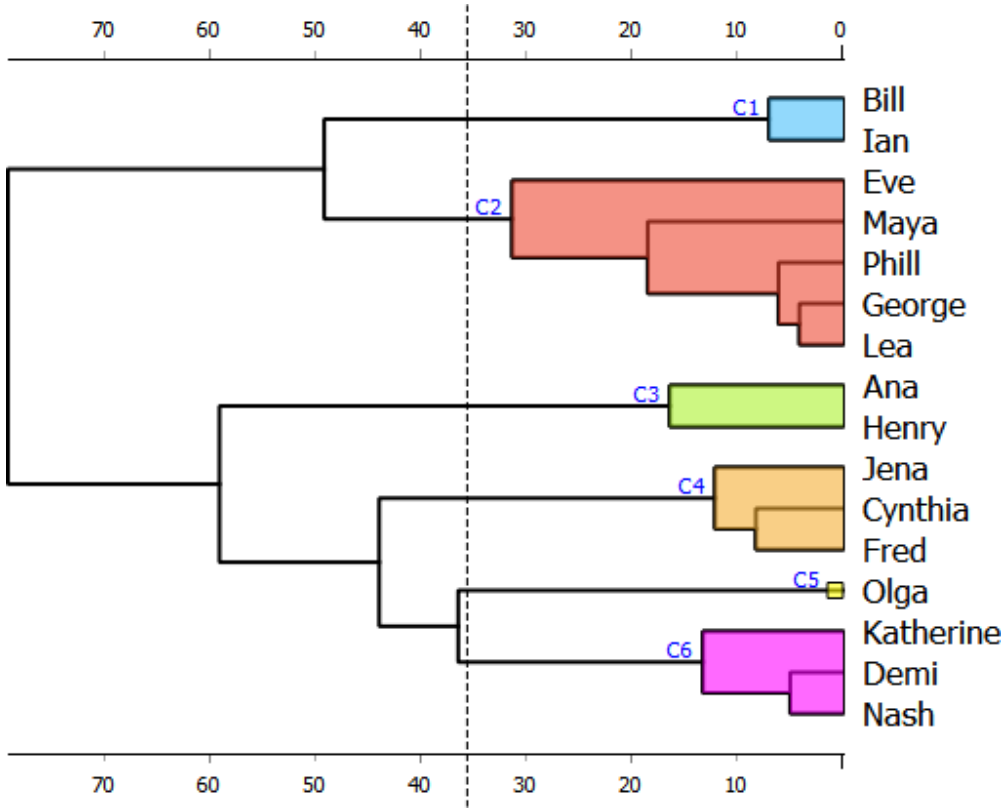
- **Ward Linkage**, auch als Ward-Minimum-Varianz-Methode bekannt, zielt darauf ab, die interne Varianz in Clustern zu minimieren.
  - Bei der Fusion von Clustern werden diejenigen zusammengeführt, deren Zusammenführung die Gesamtvarianz innerhalb des Clusters am wenigsten erhöht.
  - Ziel ist es, die Zunahme der Summe der quadrierten Abweichungen (Sum of Squared Errors, SSE) zu minimieren.
  - Das Ergebnis sind Cluster, deren Punkte nahe an ihrem Zentrum liegen, was zu einer hohen Dichtezusammenstellung führt.
  - Dieses Verfahren ist besonders nützlich, wenn es darum geht, gut abgegrenzte und kompakte Cluster zu bilden.
  - Kann als ein Kompromiss zwischen Single und Average Linkage angesehen werden, mit einer Tendenz zur Minimierung der internen Clusterstreuung.
  - Durch die Fokussierung auf die Minimierung der Varianz unterstützt Ward Linkage die Bildung homogener Gruppen, wodurch es für viele praktische Anwendungen geeignet ist.





Visualisierung mit dem Dendrogramm

- Ein Dendrogramm ist eine baumartige Diagrammstruktur, die die Bildung von Clustern visualisiert.
- Zeigt die Reihenfolge und Distanz der verschmolzenen Cluster.
- Linien im Dendrogramm kreuzen sich nicht, was die hierarchische Struktur verdeutlicht.
- Erlaubt die Betrachtung der Clusterbildung auf verschiedenen Ebenen der Hierarchie.
- Beispielhafte Anwendung auf die Noten von Schülern in Englisch und Algebra.
  - Visualisierung der schrittweisen Clusterbildung im Dendrogramm.
  - Identifikation von Schülergruppen mit ähnlichen Leistungen.
  - Möglichkeit, das Dendrogramm an verschiedenen Punkten zu "schneiden", um eine bestimmte Anzahl von Clustern zu erhalten.
- Verschieben der Schnittlinie im Dendrogramm ermöglicht die Anpassung der Clusteranzahl.



Hierarchisches Clustering: Interpretation

- Dendrogramme bieten eine detaillierte Einsicht in die Datenstruktur und mögliche Cluster.
- Die Interpretation und Entscheidung über die Anzahl der Cluster obliegt oft den Fachexpertinnen und -experten
- Hierarchisches Clustering kann für multidimensionale Daten erweitert werden.
- Visualisierung der Cluster im Streudiagramm zur Überprüfung der Gruppierung.
- Die Entscheidung über die optimale Anzahl der Cluster hängt vom Kontext und Ziel der Analyse ab.
- Hierarchisches Clustering bietet eine flexible und intuitive Methode zur Datenstrukturierung.
- Dendrogramme sind ein mächtiges Werkzeug zur Visualisierung und Analyse der Clusterbildung.
- Die Wahl der Schnitthöhe im Dendrogramm bestimmt die Anzahl und Grösse der Cluster.

Hierarchisches Clustering in höheren Dimensionen

- Mehrdimensionales Clustering ermöglicht die Analyse komplexerer Datenstrukturen.
  - Bisherige Anwendung des hierarchischen Clusterings auf zwei Dimensionen: Englisch und Algebra.
  - Einführung zusätzlicher Fächer erweitert den Datenraum auf drei, vier oder mehr Dimensionen.
  - Die Euklidische Distanz zwischen Datenpunkten wird entsprechend der Anzahl der Dimensionen erweitert.
- Die Berechnungsformel der Euklidischen Distanz skaliert mit der Anzahl der Dimensionen.
  - Jede zusätzliche Dimension fügt einen weiteren Term der quadrierten Differenz hinzu.
  - $d = \sqrt{(q_{x1} - p_{x1})^2 + (q_{x2} - p_{x2})^2 + (q_{x3} - p_{x3})^2 + \cdots}$
- Problem: Der Fluch der Dimensionalität

Der Fluch der Dimensionalität

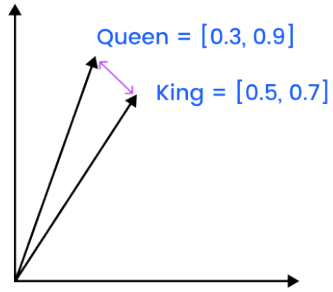
- Der «Fluch der Dimensionalität» beschreibt das Phänomen, dass mit zunehmender Anzahl an Dimensionen in einem Datensatz die Distanz zwischen den Datenpunkten immer weniger aussagekräftig wird.
  - Mit zunehmender Dimensionalität wächst der Raum exponentiell, und Datenpunkte werden «entfernter».
  - Dies liegt daran, dass der Raum so gross wird, dass die Datenpunkte im Verhältnis dazu sehr weit voneinander entfernt sind, was die Analyse und das Auffinden von Mustern erschwert.
- Schwierigkeit, in hochdimensionalen Räumen intuitive Distanzkonzepte anzuwenden.
- Kann die Effektivität von Clustering-Algorithmen beeinträchtigen und zu weniger aussagekräftigen Clustern führen.
- Bewusste Datenvorverarbeitung und -normalisierung werden wichtiger, um den Fluch der Dimensionalität zu mindern.

Der Fluch der Dimensionalität – Euklidische vs. Cosinus-Distanz (1)

- Bei der euklidischen Distanz, die die direkte Linie zwischen zwei Punkten in einem mehrdimensionalen Raum misst, wird dieser Effekt besonders deutlich.
  - In hohen Dimensionen nähern sich die euklidischen Distanzen zwischen vielen Punktepaaren einem ähnlichen Wert an, was bedeutet, dass es schwieriger wird, Unterschiede zwischen den Punkten zu erkennen.
- Die Cosinus-Distanz hingegen misst den Winkel zwischen zwei Vektoren und ist daher weniger anfällig für den Fluch der Dimensionalität.
  - Sie kann auch in hochdimensionalen Räumen nützlich sein, da sie eher die Richtung als die Länge der Vektoren berücksichtigt.
  - Dies kann besonders in Text-Mining und Information Retrieval hilfreich sein, wo es mehr um die Orientierung der Datenpunkte (z.B. Wörter in einem Dokument) als um ihre absolute Position geht.
  - Heutige Word Embedding Verfahren wie SBERT sind hochdimensional

Der Fluch der Dimensionalität – Euklidische vs. Cosinus-Distanz (2)

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

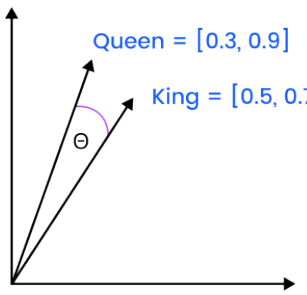


Vector Q = [0.3, 0.9]

Vector K = [0.5, 0.7]

$$d(Q,K) = \sqrt{(0.3-0.5)^2 + (0.9-0.7)^2}$$
$$= \sqrt{(0.2)^2 + (0.2)^2}$$
$$= \sqrt{0.04 + 0.04}$$
$$0.28 \cong \sqrt{0.08}$$

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$



V ct Q = [0.3, 0.9]

V ct K = [0.5, 0.7]

$$cos(Q,K) = \frac{(0.3 \times 0.5) + (0.9 \times 0.7)}{\sqrt{0.3^2 + 0.9^2} \times \sqrt{0.5^2 + 0.7^2}}$$
$$= \frac{0.15 + 0.63}{\sqrt{0.9} \times \sqrt{0.74}}$$
$$0.96 \cong \frac{0.78}{\sqrt{0.9} \times \sqrt{0.74}}$$

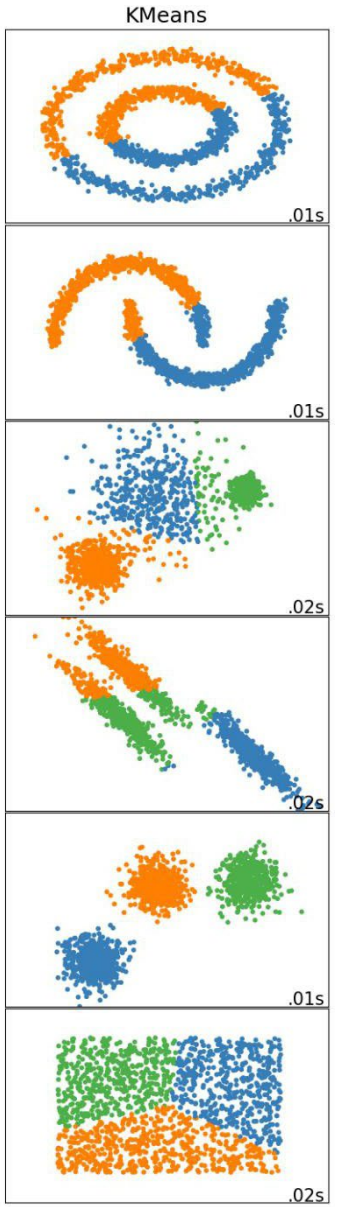
$$d(Q,K) = 1 - \cos(Q,K)$$
$$0.04 \cong 1 - 0.96$$

Beispiele: https://zilliz.com/blog/similarity-metrics-for-vector-search

k-Means Clustering

k-Means Clustering

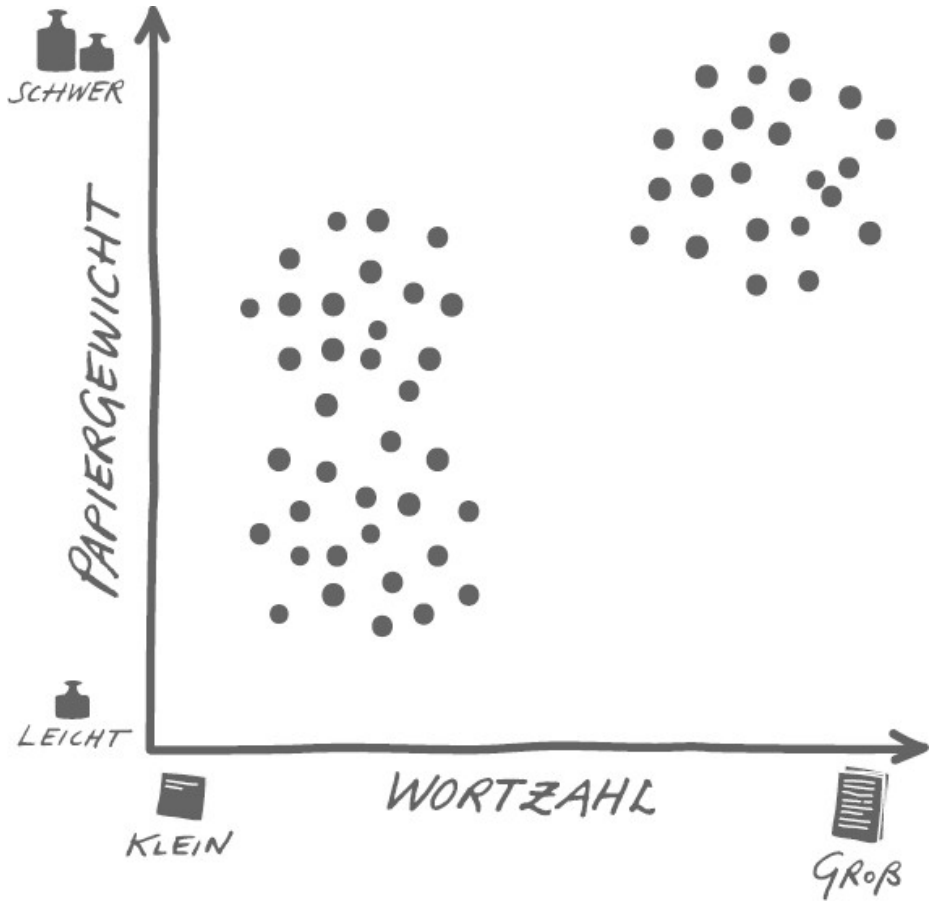
- k-Means ist ein partitiver Clustering-Algorithmus, der Datenpunkte in eine vorbestimmte Anzahl von Clustern teilt.
  - Ziel ist es, Datenpunkte in k vordefinierte Gruppen (Cluster) zu teilen.
- Anwendbar bei verschiedenen Datentypen, um Muster zu erkennen.
- Benötigt die Angabe der Clusteranzahl (k) vor der Ausführung.
- Nutzt Distanzmetriken, um die Ähnlichkeit zwischen Datenpunkten zu bestimmen.
- Iterativer Ansatz: Beginn mit zufälligen Mittelpunkten, gefolgt von Neuzuordnung der Datenpunkte und Aktualisierung der Mittelpunkte.
- Eignet sich zur Gruppierung ähnlicher Objekte ohne vorherige Kenntnis ihrer Kategorien.



Müller, D. (2019). k-Means-Algorithmus. In K. Kersting, C. Lampert, & C. Rothkopf (Eds.), Wie Maschinen lernen (pp. 81–88). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26763-6_11

Vorbereitung der Daten für das Clustering – Beispiel

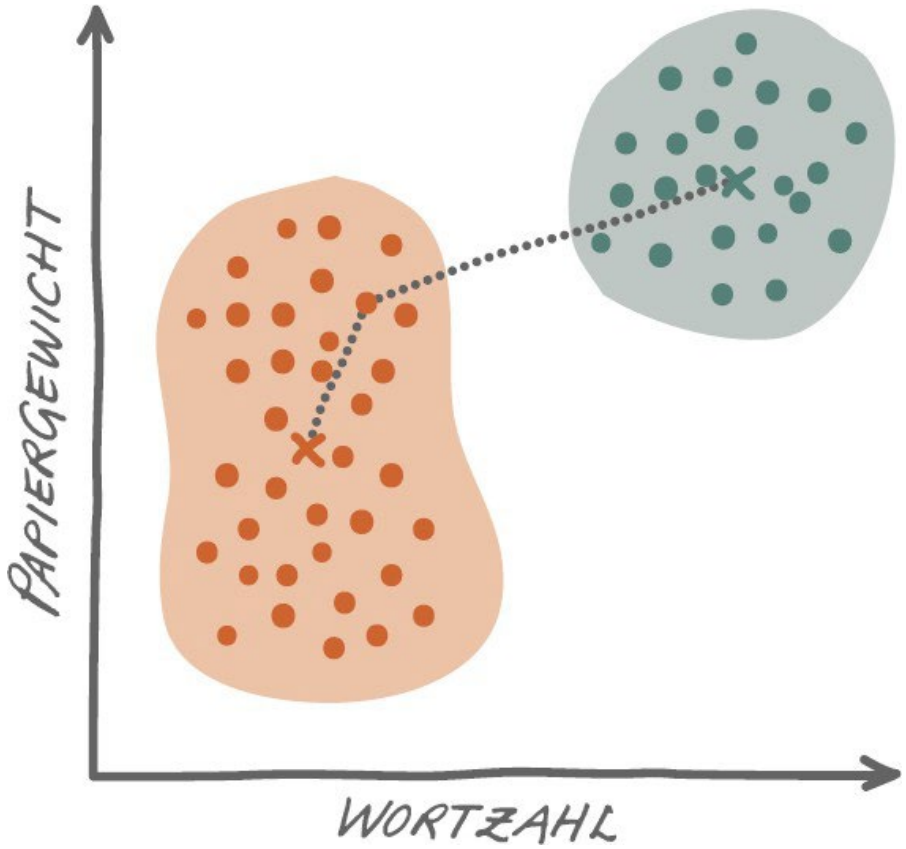
- Erhebung und Dokumentation der Merkmale für jedes Objekt, in diesem Beispiel: Dokumente.
  - Zwei gewählte Eigenschaften für die Clustering-Analyse: Länge der Dokumente (Anzahl der Wörter) und Papiergewicht.
- Diese Eigenschaften werden in einem Koordinatensystem visualisiert, wobei jeder Punkt ein Dokument repräsentiert.
- Ziel ist es, anhand dieser Merkmale Ähnlichkeiten zwischen den Dokumenten zu erkennen und sie entsprechend zu gruppieren.
- Die Wahl der Merkmale ist entscheidend für die Effektivität des Clustering-Prozesses.
- Diese Vorbereitungsschritte sind grundlegend für die Anwendung des k-Means-Algorithmus.



Müller, D. (2019). k-Means-Algorithmus. In K. Kersting, C. Lampert, & C. Rothkopf (Eds.), Wie Maschinen lernen (pp. 81–88). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26763-6_11

Ziel: Bestimmung der Clusterzugehörigkeit durch Distanzmessung

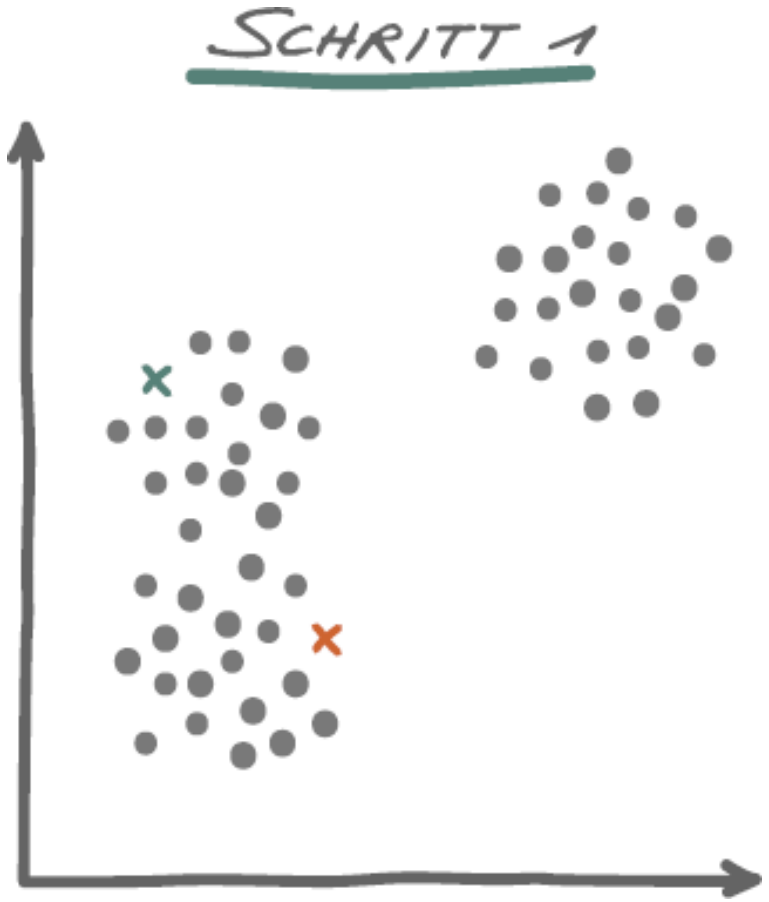
- Erläuterung der Rolle der Distanzmessung im Clustering-Prozess.
- Die Distanz eines Punktes (Dokument) zu den Mittelpunkten der Cluster ist entscheidend für die Zuordnung.
- Kürzeste Distanz bestimmt, zu welchem Cluster ein Dokument gehört.
- Visualisierung der Distanzen hilft bei der Veranschaulichung, wie der Algorithmus Gruppenzugehörigkeiten entscheidet.
- Die Berechnung und Aktualisierung der Mittelpunkte basieren auf dieser Distanzmessung.
- Distanzmessungen sind ein Kernbestandteil der iterativen Verbesserung der Clusterbildung.
- Die Abbildung verdeutlicht, dass die Optimierung der Mittelpunkte eine zentrale Aufgabe des k-Means-Algorithmus ist.



Müller, D. (2019). k-Means-Algorithmus. In K. Kersting, C. Lampert, & C. Rothkopf (Eds.), Wie Maschinen lernen (pp. 81–88). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26763-6_11

k-Means-Algorithmus – Schritt 1: Initialisierung

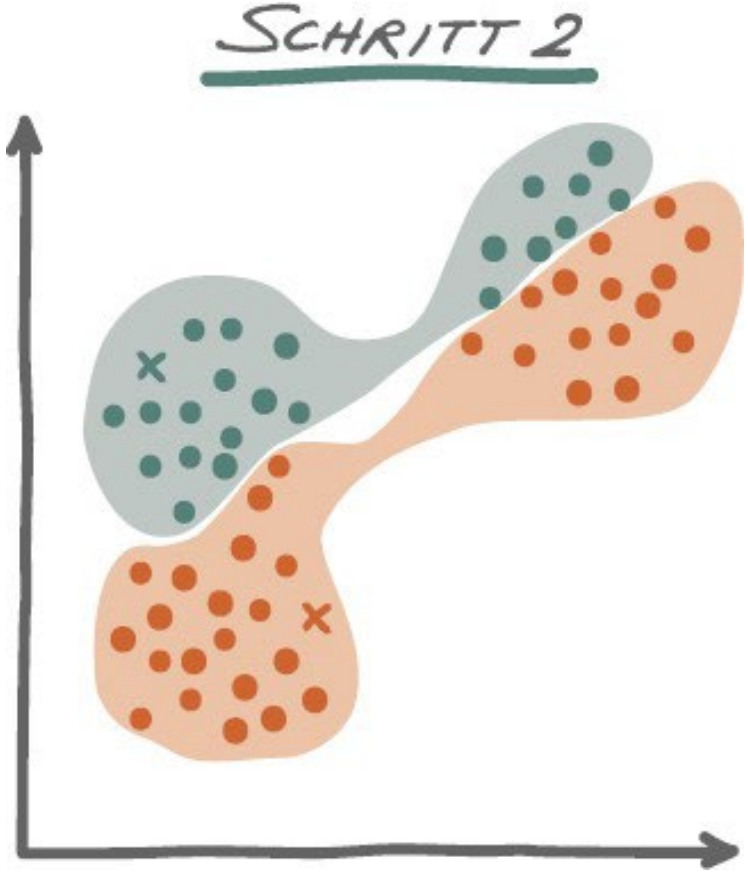
- Auswahl von k zufälligen Datenpunkten als anfängliche Cluster-Mittelpunkte.
- Diese Anfangswahl kann das Ergebnis und die Konvergenzgeschwindigkeit beeinflussen.
- Initialisierungsmethoden variieren zur Optimierung der Ergebnisse.
- Die Wahl der Mittelpunkte ist entscheidend für die folgende Clusterbildung.
- Verschiedene Techniken existieren, um optimale Startpunkte zu bestimmen.
- Zufällige Auswahl dient der Vereinfachung und initialen Annäherung.



Müller, D. (2019). k-Means-Algorithmus. In K. Kersting, C. Lampert, & C. Rothkopf (Eds.), Wie Maschinen lernen (pp. 81–88). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26763-6_11

k-Means-Algorithmus – Schritt 2: Zuordnung zu Clustern

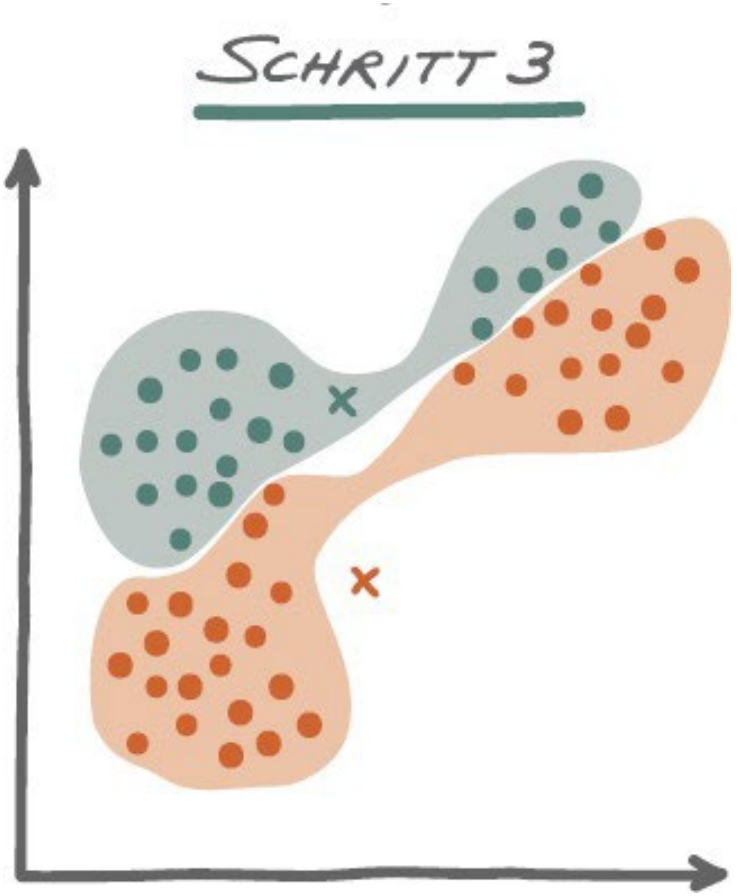
- Jeder Datenpunkt wird dem nächstgelegenen Cluster-Mittelpunkt zugeordnet.
- Distanzmasse (z.B. Euklidische Distanz) bestimmen die Nähe zu Mittelpunkten.
- Ergebnis ist eine vorläufige Gruppierung der Daten basierend auf der aktuellen Mittelpunktlage.
- Diese Zuordnung bildet die Grundlage für die Aktualisierung der Mittelpunkte.
- Die Zuordnung wird in jedem Iterationsschritt neu berechnet.
- Flexibilität der Zuordnung ermöglicht die Anpassung und Optimierung der Cluster.



Müller, D. (2019). k-Means-Algorithmus. In K. Kersting, C. Lampert, & C. Rothkopf (Eds.), Wie Maschinen lernen (pp. 81–88). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26763-6_11

k-Means-Algorithmus – Schritt 3: Aktualisierung der Cluster-Mittelpunkte

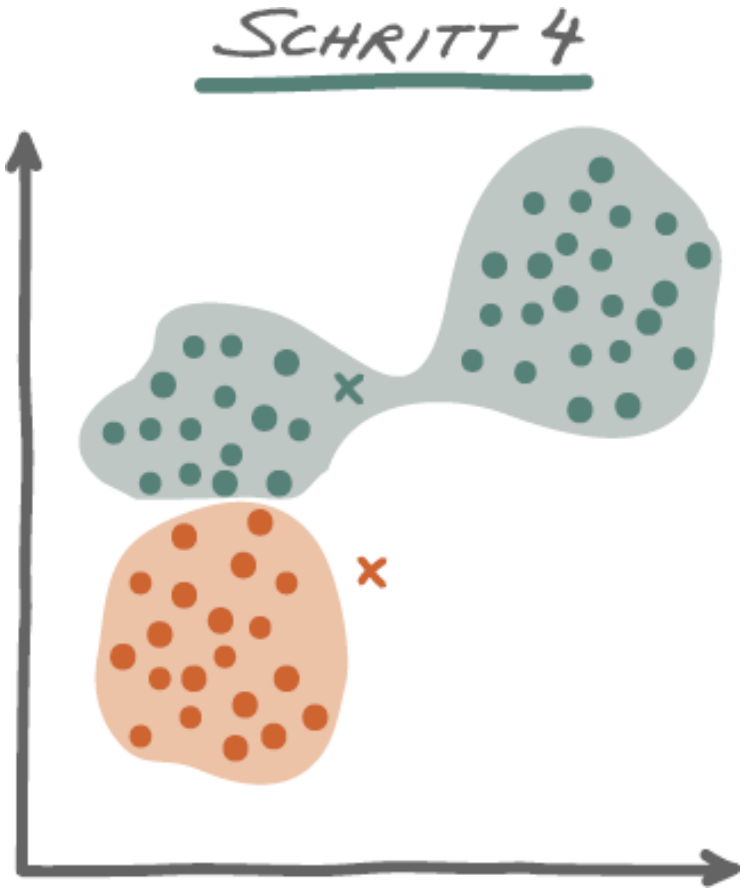
- Neuberechnung der Mittelpunkte als Durchschnitt aller zugeordneten Datenpunkte.
- Ziel ist die Minimierung der inneren Cluster-Variabilität.
- Die Position der neuen Mittelpunkte kann sich deutlich von der ursprünglichen unterscheiden.
- Aktualisierung verbessert die Repräsentativität der Mittelpunkte für ihre Cluster.
- Dieser Schritt erhöht die Homogenität innerhalb der Gruppen.
- Iterationen führen zur Feinabstimmung der Clusterformation.



Müller, D. (2019). k-Means-Algorithmus. In K. Kersting, C. Lampert, & C. Rothkopf (Eds.), Wie Maschinen lernen (pp. 81–88). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26763-6_11

k-Means-Algorithmus – Schritt 4: Wiederholung und Konvergenzprüfung

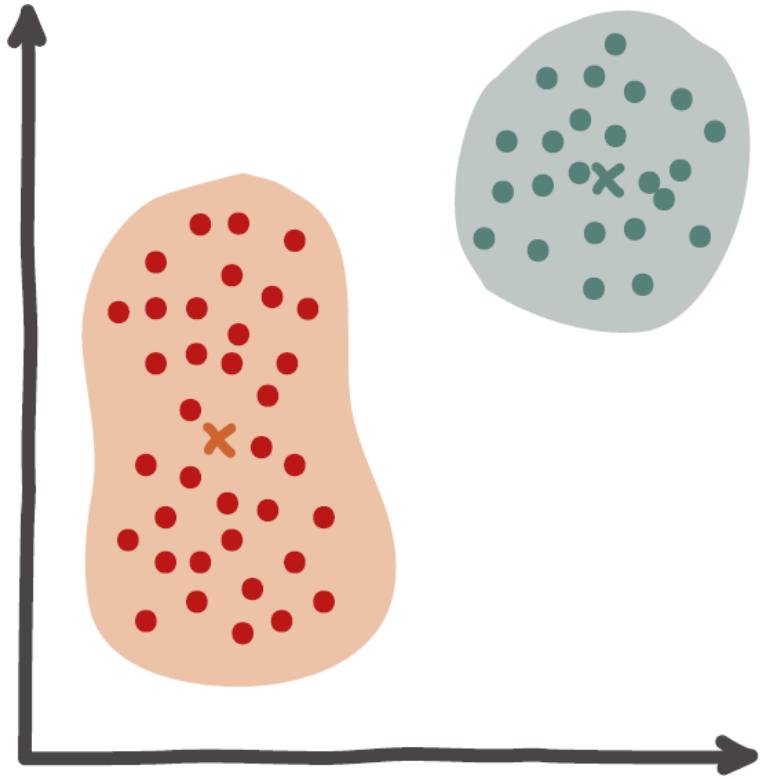
- Wiederholung der Schritte 2 und 3, bis die Mittelpunkte stabil bleiben.
  - Ein Cluster ist optimal, wenn keine Datenpunkte mehr ihre Zugehörigkeit ändern (Schritt 4).
- Konvergenz bedeutet, dass sich die Mittelpunkte und Clusterzugehörigkeiten nicht mehr verändern.
- Das Verfahren endet, wenn die Cluster stabil sind oder eine maximale Anzahl von Iterationen erreicht ist.
- Die Iteration verbessert die Genauigkeit und Relevanz der gebildeten Cluster.
- Manchmal sind mehrere Durchläufe mit verschiedenen Startwerten nötig.



Müller, D. (2019). k-Means-Algorithmus. In K. Kersting, C. Lampert, & C. Rothkopf (Eds.), Wie Maschinen lernen (pp. 81–88). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26763-6_11

k-Means-Algorithmus – Eigenschaften und Konvergenz

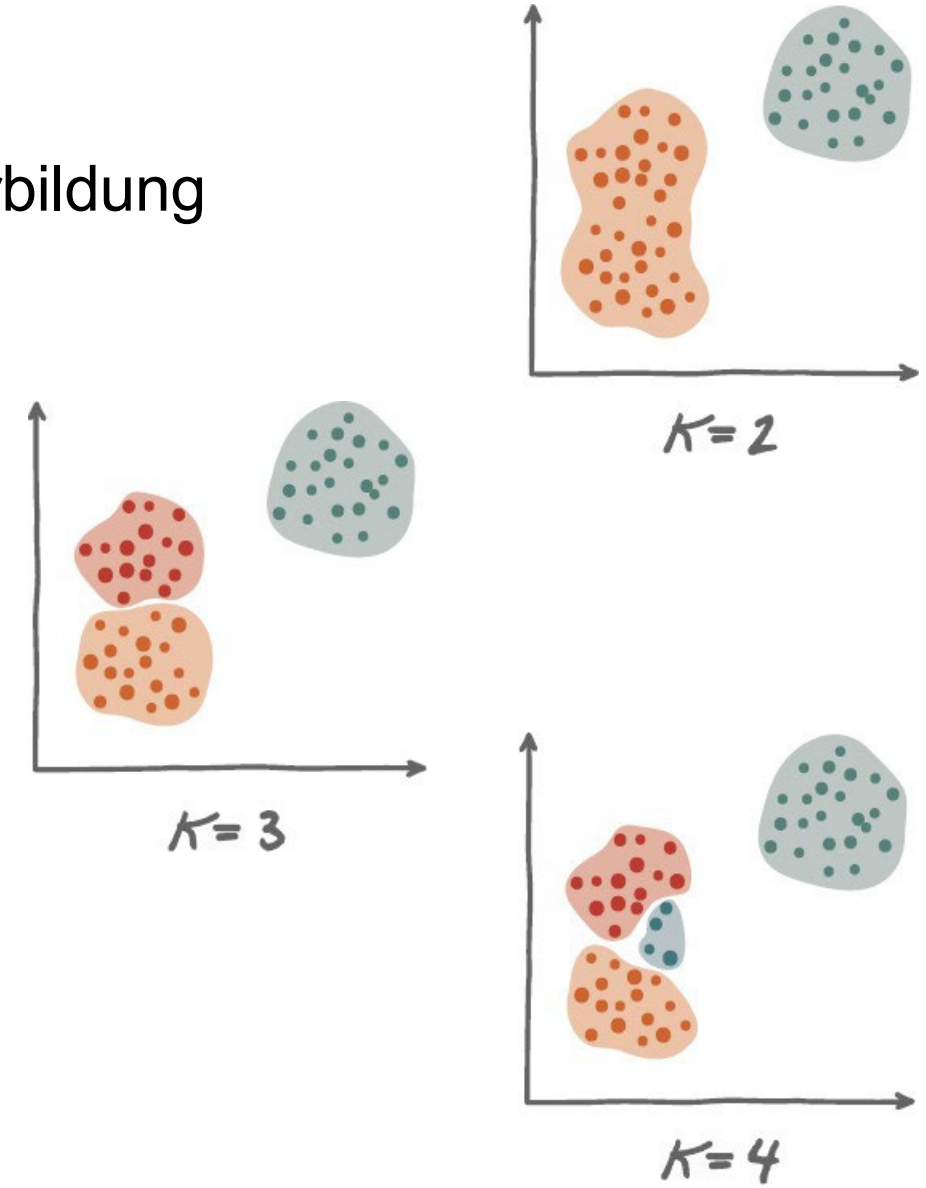
- Konvergenz ist erreicht, wenn sich die Mittelpunkte nicht mehr signifikant verschieben und die Zuordnungen stabil bleiben.
- Das Kriterium für Konvergenz kann auch eine vorher festgelegte maximale Anzahl von Iterationen sein.
- Stabile Cluster bedeuten, dass weitere Iterationen keine signifikante Verbesserung der Clusterzuordnungen bewirken.
- Manchmal können verschiedene Ausführungen des Algorithmus mit unterschiedlichen Anfangsmittelpunkten hilfreich sein, um die Robustheit der gefundenen Cluster zu überprüfen.
- Die Endphase des Algorithmus bietet eine Gruppierung, die eine fundierte Analyse und Interpretation der Daten ermöglicht.



Müller, D. (2019). k-Means-Algorithmus. In K. Kersting, C. Lampert, & C. Rothkopf (Eds.), Wie Maschinen lernen (pp. 81–88). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26763-6_11

Einfluss der Anzahl der Cluster (k) auf die Clusterbildung

- Mit k = 2 bildet der Algorithmus zwei grundlegende Gruppen, basierend auf den ermittelten Merkmalen.
- Erhöhung auf k = 3 führt zur Identifikation einer zusätzlichen, differenzierten Gruppe innerhalb der Daten.
- Bei k = 4 werden die Daten in noch spezifischere Gruppen unterteilt, die feinere Unterschiede hervorheben.
- Die Wahl von k hat signifikanten Einfluss auf die Granularität und Interpretierbarkeit der Cluster.
- Zu viele Cluster können zur Übersegmentierung führen, während zu wenige wichtige Unterschiede verbergen können.
- Die optimale Anzahl von Clustern (k) zu bestimmen, bleibt eine zentrale Herausforderung im Clustering-Prozess.



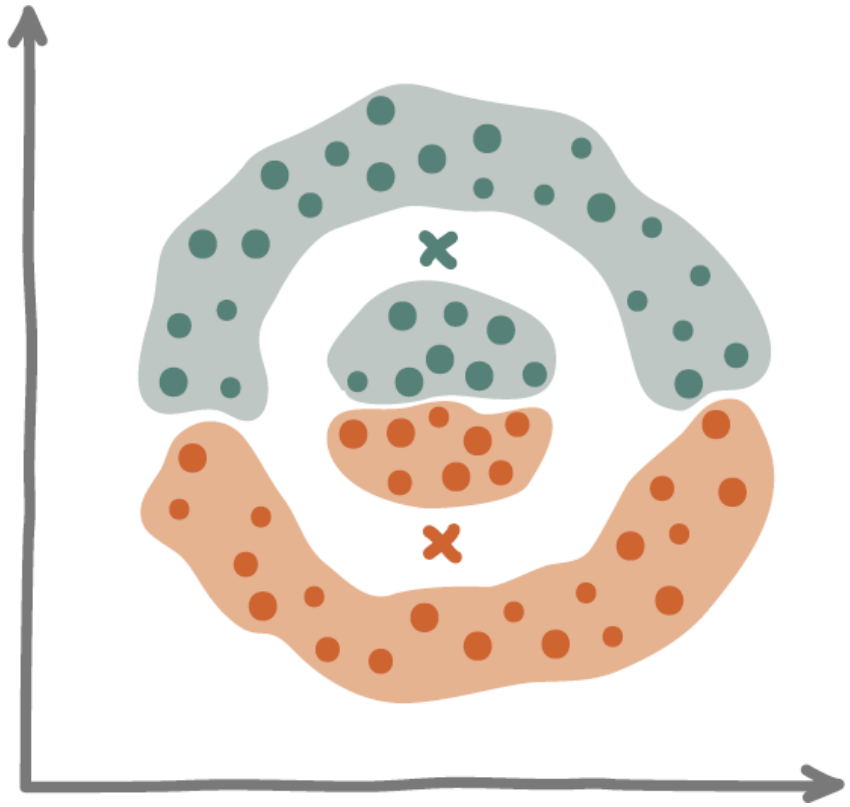
Müller, D. (2019). k-Means-Algorithmus. In K. Kersting, C. Lampert, & C. Rothkopf (Eds.), Wie Maschinen lernen (pp. 81–88). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26763-6_11

Vorteile von k-Means gegenüber Hierarchischem Clustering

- Skalierbarkeit: Effizient bei grossen Datensätzen.
- Speicheranforderungen: Benötigt nicht die Speicherung einer kompletten Distanzmatrix, was im Vergleich zum hierarchischen Clustering erheblich Speicherplatz spart.
- Geschwindigkeit: Schnellere Konvergenz, oft nur nach wenigen Iterationen, während hierarchisches Clustering bei grossen Datenmengen praktisch unanwendbar sein kann.
- Flexibilität in der Cluster-Form: Kann effektiv runde (sphärische) Cluster identifizieren, was für viele praktische Anwendungen ausreichend ist.
- Einfache Anpassung: Ermöglicht einfache Manipulation durch Vorwählen der Clusteranzahl und gegebenenfalls Neuplatzierung der Zentroide (Mittelpunkten).

Herausforderungen bei problematischen Datenstrukturen

- Auswahl der optimalen Anzahl von Clustern (k) ist entscheidend und nicht immer offensichtlich.
- Sensibilität gegenüber den Startpositionen der Mittelpunkte und deren Einfluss auf das Endergebnis.
- Problematische Daten können z.B. durch Ausreisser, Überlappungen oder nicht-kugelförmige Cluster gekennzeichnet sein.
  - Ausreisser können die Berechnung der Cluster-Mittelpunkte beeinträchtigen und die Clusterzuordnung verzerren.
  - Überlappende Cluster können zu Unsicherheiten bei der Zuordnung von Datenpunkten zu den nächstliegenden Clustern führen.
  - Nicht-kugelförmige Cluster stellen eine Herausforderung dar, da k-Means auf Distanzmessungen basiert, die kugelförmige Cluster bevorzugen.
- Eine sorgfältige Vorverarbeitung der Daten und die Wahl geeigneter Merkmale sind essenziell für die Verbesserung der Clustering-Ergebnisse.

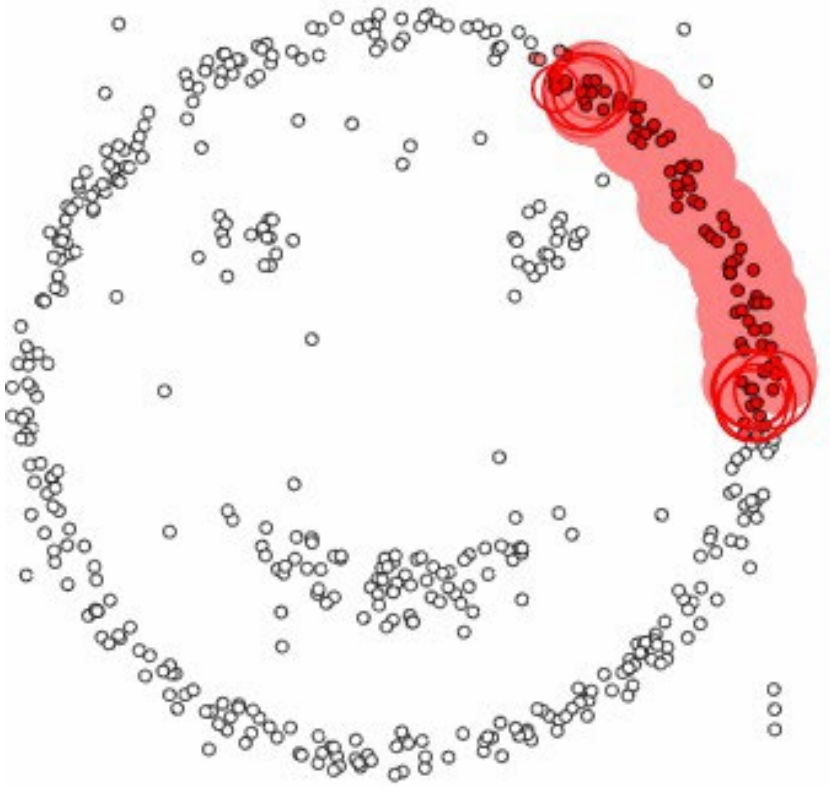


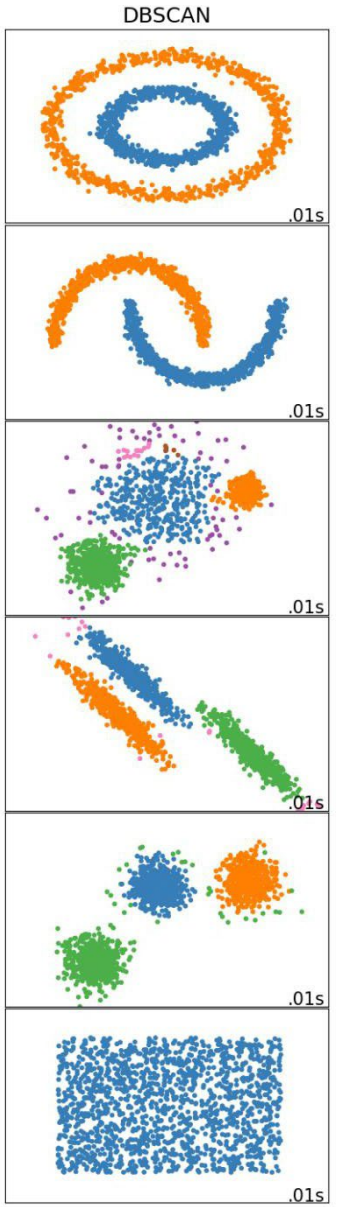
Müller, D. (2019). k-Means-Algorithmus. In K. Kersting, C. Lampert, & C. Rothkopf (Eds.), Wie Maschinen lernen (pp. 81–88). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-26763-6_11

DBSCAN Clustering

DBSCAN (1)

- DBSCAN steht für "Density-Based Spatial Clustering of Applications with Noise".
- Es identifiziert Cluster basierend auf der Dichte der Datenpunkte in einem Raum.
  - Es geht um die räumliche Nähe von Datenpunkten!
- Kann effektiv mit Ausreissern umgehen und erfordert keine Vorgabe der Clusteranzahl.
- Geeignet für Daten mit komplexen Strukturen und unterschiedlichen Dichten.





Animation: https://www.digitalvidya.com/blog/the-top-5-clustering-algorithms-data-scientists-should-know/

DBSCAN (2)

- Kernpunkte: Punkte, die mindestens eine Mindestanzahl von Nachbarn in einem gegebenen Radius haben.
- Randpunkte: Punkte, die weniger Nachbarn als Kernpunkte haben, aber in der Nähe von Kernpunkten liegen.
- Rauschen: Punkte, die weder Kern- noch Randpunkte sind.
- Der Algorithmus lässt Cluster durch Verknüpfen von Kernpunkten wachsen, die innerhalb des definierten Radius voneinander entfernt sind.

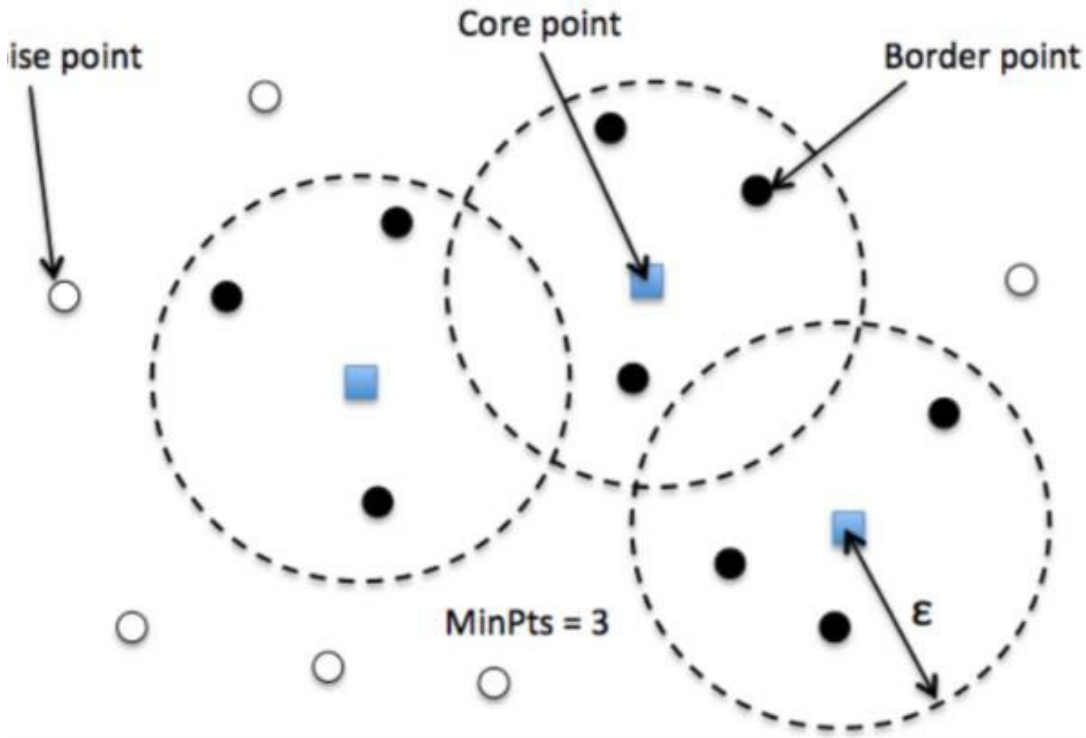


Bild: https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html

Vergleich von DBSCAN, k-Means und Hierarchischem Clustering

- DBSCAN: Ideal für komplexe Datensätze mit variabler Clusterdichte und für Anwendungen, bei denen Ausreisser präsent sind.
  - Keine Notwendigkeit, die Anzahl der Cluster vorzugeben.
  - Kann Cluster unterschiedlicher Formen und Grössen erkennen.
  - Effektiv in der Behandlung von Rauschen und Ausreissern.
- k-Means: Beste Wahl für grosse, klar abgegrenzte und sphärische Cluster, schnelle Verarbeitung erforderlich.
  - Schnell und effizient bei grossen Datensätzen mit sphärischen Clustern.
  - Erfordert die Vorgabe der Anzahl der Cluster.
  - Empfindlich gegenüber Ausreissern und Anfangsplatzierung der Zentroide.
- Hierarchisches Clustering: Empfohlen für explorative Datenanalyse, wenn visuelle Darstellungen der Clusterhierarchie wertvoll sind.
  - Ermöglicht eine detaillierte Darstellung der Datenstruktur durch ein Dendrogramm.
  - Nicht skalierbar für sehr grosse Datensätze.
  - Geeignet für kleinere Datensätze und wenn die Clusteranzahl unbekannt ist.

Word Clustering

Clustering Anwendung – Word Clustering

- Wörter können in einem Vektorraum eingebettet werden, wobei ähnliche Wörter ähnliche Vektoren haben.
- Ziel ist es, semantisch ähnliche Wörter zu identifizieren und zu gruppieren.
- Word Embeddings wie fastText wandeln Wörter in Vektoren mit mehreren Dimensionen um.
  - Jedes Wort wird durch einen Vektor mit 300 Merkmalen repräsentiert.
- Erstellung einer Distanzmatrix, die die semantischen Abstände zwischen allen Wörtern zeigt.
  - Nutzung des Cosine-Distanzmasses zur Messung der Ähnlichkeit zwischen Wörtern.
- Einsatz von hierarchischem Clustering, um ein Dendrogramm der Wortbeziehungen zu erzeugen.
  - Identifikation von thematischen Clustern wie (siehe Beispieldatensatz) Musikinstrumente, Schreibwerkzeuge und Nahrungsmittel.

Word Clustering – Visualisierung mit t-SNE

- t-SNE hilft dabei, hochdimensionale Daten auf eine Weise zu reduzieren, dass ähnliche Punkte (Wörter) nahe beieinander liegen.
- Jeder Punkt im t-SNE-Plot repräsentiert ein Wort.
- Selektion und Anzeige spezifischer Wortgruppen im t-SNE-Plot zur visuellen Überprüfung der Clusterkonsistenz.
- Hierarchisches Clustering und t-SNE bieten verschiedene Einblicke und können kombiniert werden, um ein tieferes Verständnis der Datenstruktur zu erreichen.
- Hierarchisches Clustering und t-SNE sind wertvolle Werkzeuge für die explorative Datenanalyse von Wortbedeutungen.
  - Überprüfung, ob die im Dendrogramm (Hierarchisches Clustering) identifizierten Cluster mit den Gruppen im t-SNE-Plot übereinstimmen.
  - Möglichkeit, Inkonsistenzen zu entdecken, wie das Wort «whistle» (siehe Beispieldatensatz), das im t-SNE-Plot isoliert erscheint.