Unsupervised Learning

Clustering	"Ich sortiere Dinge so, dass Ähnliche zusammengehören." Wie wenn du Spielzeug sortierst: • Alle Legosteine zusammen • Alle Kuscheltiere zusammen • Alle Autos zusammen
Dimensionsreduktion	Stell dir vor, du beschreibst eine Blume mit 4 Eigenschaften: • Wie **lang** ist das Blatt? • Wie **breit** ist das Blatt? • Wie **lang** ist das Blütenblatt? • Wie **breit** ist das Blütenblatt? Das sind 4 Infos – man sagt auch: **4 Dimensionen**. **Dimensionsreduktion bedeutet**: "Ich versuche, das Wichtigste aus den 4 Eigenschaften in nur 2 oder 3 zusammenzufassen." So wie wenn du sagst: "Ich muss nicht alles wissen, es reicht, wenn ich nur das Wichtigste sehe."
Assoziation Rule Mining	Das ist ein Verfahren, mit dem der Computer herausfindet: **Welche Produkte kaufen Menschen oft zusammen?** Wenn jemand **Brot** kauft, dann kauft er vielleicht auch oft **Butter** dazu.
Outlier Detection / Auffälligkeiten erkennen	Der Computer schaut Daten an und findet Datenpunkte, die ganz anders sind als der Rest. *„Finde die Daten, die nicht normal aussehen!“* Damit man schnell reagieren kann!
Früher Feature-based Machine Learning	• Der Mensch musste **wichtige Merkmale** (z. B. Formen, Kanten) **von Hand rausziehen** (mithelfen) • Dann wurde damit ein Algorithmus trainiert (z. B. Entscheidungsbaum) • Beispiel: "Oh, das hat Streifen, vielleicht ein Tiger!"
Heute Deep Learning	**Heute (Deep Learning):** • Der Computer **lernt die wichtigen Merkmale selbst** • Er schaut sich **direkt die Pixel** an und erkennt selbst, was wichtig ist. • Das nennt man **Neuronales Netzwerk**
Clusteranalyse	Clusteranalyse ist wie **"automatisch Gruppen bilden",** der Computer sortiert Leute oder Dinge **nach Ähnlichkeiten** in Gruppen ein. *„Clusteranalyse findet heraus, wer ähnlich tickt – ganz ohne vorher zu wissen, wer zu wem gehört.“*
Ziele Clusteranalyse	• Ähnliche Dinge in Gruppen einteilen (z. B. Personen, Produkte). • Man weiss **vorher nicht**, wie viele Gruppen es gibt – das findet der Computer raus. • Wird z. B. für **Marketing oder soziale Netzwerke** genutzt.
Herausforderungen Clusteranalyse	• Wie viele Gruppen soll man machen? • Manche Daten passen **nicht klar** in eine Gruppe.
Nutzen	• Hilft, Daten besser zu verstehen und **klügere Entscheidungen** zu treffen.
Euklidische Distanz	Wie weit sind zwei Punkte voneinander entfernt? - Hilft Gruppen zu bilden (k-means) - Nützlich für: Nutzergruppen, Kaufverhalten, Medizin Muster - Macht Datenmengen übersichtlicher

Datenauswahl & Vereinfachung	„Die Auswahl repräsentativer Merkmale ist entscheidend für die Effektivität des Clusterings.
Visualisierung im Streudiagramm (Scatter Plot)	- ermöglichen intuitive Einschätzung Datenstrutkur - Label erleichtern Identifikation von Ausreissern & Mustern - Liefern Anhaltspunkte für Bildung Cluster - Visualisierung mächtiges Werkzeug Hypothesenbildung
Euklidische Distanz Berechnung	Misst direkten Abstand zwischen zwei Punkten $Euclidean\ Dinstance\ (d) = \sqrt{(q_{x1} - p_{x1})^2 + (q_{x2} - p_{x2})^2}$
Anwendung Euklidischen Distanz Clustering	- Bildet Fundament für quantitative Analyse Ähnlichkeiten - Erleichtert Dateninterpretation durch Bildung Gruppen - Anwendbar Kundenanalyse
Hierarchisches Clustering	- Organisiert Datenpunkte Baumstruktur - Jeder Datenpunkt eigenständiges Cluster - Führt schrittweise nahe Cluster zusammen - Wahl Distanzmasses beeinflusst Clusterstrukturkannst
Distanzmasse Single Linkage (Nächster Nachbar)	- Schaut sich den kürzesten Abstand zwischen zwei Punkten aus unterschiedlichen Gruppen an. - Wenn zwei Punkte sehr nah sind, verbindet es die Gruppen. - Kann dabei lange „Ketten“ bilden – auch wenn der Rest gar nicht so nah beieinander ist. - Empfindlich bei Ausreissern. Gut, wenn man nicht-kreisförmige Gruppen erkennen will.
Distanzmasse Average Linkage (Durchschnitt)	- Rechnet den durchschnittlichen Abstand zwischen allen Punkten zweier Gruppen. - Dadurch entstehen sauberere Gruppen – eher kompakt und gut getrennt. - Weniger empfindlich gegenüber Ausreissern. - Gibt oft bessere, gleichmässigere Ergebnisse.
Distanzmasse Ward Linkage	Ward Linkage sorgt dafür, dass die **Gruppen möglichst "sauber"** bleiben. Das heißt: Die Punkte in einem Cluster sollen **nah beieinander liegen** (wenig Streuung). **Wie funktioniert das?** • Es schaut: **Welche zwei Gruppen kann ich verbinden**, ohne dass die Punkte zu unterschiedlich werden? • Ziel: Die Punkte in einem Cluster **sollen sich möglichst ähnlich sein**. **Vorteile:** • Macht **runde, kompakte Gruppen**, die gut getrennt sind. • Funktioniert gut für viele reale Anwendungen. • Ist wie ein **Kompromiss** zwischen "nächster Punkt" (Single Linkage) und "Durchschnitt" (Average Linkage).
Visualisierung Dendrogramm	Baumartige Diagrammstruktur, die die Bildung von Clustern visualisiert. Man kann gut sehen welche Dinge zusammengehören.
Hierarchisches Clustering Interpretation	- Detaillierte Einsicht in Datenstruktur & mögliche Cluster - Wie viele Cluster entscheidet «Experte»¨ - Kann multidimensionale Daten erweitern - Visualisierung Cluster im Streudiagramm - Dendrogramm mächtiges Werkzeug Visualisierung

Hierarchisches Clustering in höheren Dimensionen	Mehrdimensionales Clusteirng ermöglicht Analyse komplexerer Daten. Die Abstände (Distanzen) zwischen Datenpunkten werden mit einer erweiterten Formel berechnet, je mehr Dimensionen desto länger Formel. **Problem:** zu viele Dimensionen machen die Analyse schwieriger. Das nennt man den Fluch der Dimensionalität. $d = \sqrt{(q_{x1} - p_{x1})^2 + (q_{x2} - p_{x2})^2 + (q_{x3} - p_{x3})^2 + \cdots}$
Fluch der Dimensionalität	- Je mehr Eigenschaften (Dimensionen) Daten haben, desto grösser wird Raum. - Abstände zwischen Datenpunkten sagen weniger aus - Muster zu finden schwieriger - Clustering funktioniert schlechter - Lösung: Daten gut vorbereiten und normalisieren
Euklidisch vs. Cosunus-Distanz (bei vielen Dimensionen)	**Euklidische Distanz:** misst den direkten Abstand zwischen zwei Punkten, in hohen Dimensionen alle Abstände ähnlich. **Consinus-Distanz:** misst Winkel zwischen zwei Punkten (Vektoren), funktioniert besser in vielen Dimensionen, da sie Richtung statt Abstand betrachtet, nützlich bei Textdaten.
k-Means Clustering	- Teilt Daten in vorgegebene Anzahl k an Gruppen - Ziel: ähnliche Datenpunkte in eine Gruppe zu packen - Braucht k als Startwert (wie viele Gruppen?) - Gut wenn keine Vorkenntnisse über Datenkategorien hat
k-Means Algorithmus Schritt 1 Initialisierung	- Algorithmus startet mit zufälligen Mittelpunkten - Startpunkte beeinflussen Endergebnis & wie schnell Algorithmus konvergiert - Wahl Mittelpunkte wichtig
Schritt 2 Zuordnung Clustern	- Jeder Punkt wird nächsten Mittelpunkt zugeordnet - Dafür verwendet man euklidische Distanz - So entsteht erste Gruppierung Daten - Basis für Berechnung Mittelpunkte - Wird in jedem Schritt neu berechnet, um Cluster verbessern
Schritt 3 Mittelpunkte aktualisieren	- Neuen Mittelpunkte werden als Durchschnitt Punkte in jedem Cluster berechnet - Ziel: Cluster verbessern Gruppen kleiner machen - Mittelpunkte verschieben sich, um besser zu passen - Macht Cluster einheitlicher (homogener) - Schritt wird mehrmals wiederholt bis nichts mehr ändert
Schritt 4 Wiederholen & Prüfen	- Schritt 2 & 3 wiederholt sich bis nichts mehr verändert - Dann gilt: Cluster sind stabil = Konvergenz erreicht - Algorithmus stoppt wenn alles stabil ist oder eine maximale Anzahl Wiederholungen erreicht wurde - Wiederholungen machen Cluster genauer und besser - Manchmal mehrere Durchläufe mit anderen Startpunkten nötig für gutes Ergebnis
Einfluss k (Anzahl der Cluster)	- K = 2 (Daten in 2 grosse Gruppen aufteilen) - K = 4 (Daten in 4 feinere Gruppen aufteilen) Zu viele Cluster = zu kleinteilig Zu wenige Cluster = wichtige Unterschiede gehen verloren Richtige Anzahl Cluster finden entscheidend

Vorteile k-Means gegenüber hierarchischem Clustering	- Schneller & effizienter bei grossen Datenmengen - Weniger Speicherbedarf - Schnelle Ergebnisse - Gut für runde (sphärische) Gruppen - Einfach anpassbar
Herausforderungen schwierigen Daten	Richtige Anzahl Cluster (K) nicht eindeutig, ergebnis abhängig von Startpunkten & Probleme entstehen, wenn: - Ausreisser - Cluster sich überlappen - Cluster keine runde Form haben Dann kann k-Means falsche Gruppen bilden
DBSCAN	Desity-Based Spatial Clustering of Applications with Noise - Findet Cluster, wo viele Punkte beieinander liegen - Keine Angabe Clusteranzahl nötig - Erkennt Ausreisser - Gut für komplizierte Formen & unterschiedliche Dichten in Daten
DBSCAN wichtige Punktarten	**Kernpunkte:** haben genug Nachbarn & bilden Zentrum Cluster **Randpunkte:** wenige Nachbarn, sind aber nahe bei Kernpunkten **Rauschen:** Punkte die weit weg sind, gehören zu keinem Clustern
Word Clustering	- Wörter werden als Zahlen-Vektoren dargestellt (Word Embeddings). - Ähnliche Wörter haben ähnliche Vektoren - Ziel: Thematisch ähnliche Wörter gruppieren - Dafür wird Cosinus-Distanz zur Messung ähnlichkeit genutzt - Hierarchisches Clustering macht daraus Dendrogramm
Word Clustering mit t-SNE	- **t-SNE** zeigt Wörter, dass ähnliche **Wörter beieinander** - jeder **t-SNE-Plot** ist ein Wort - so kann man Cluster visuell prüfen - **t-SNE** kann ungewöhnliche Wörter sichtbar machen, die nicht richtig passen