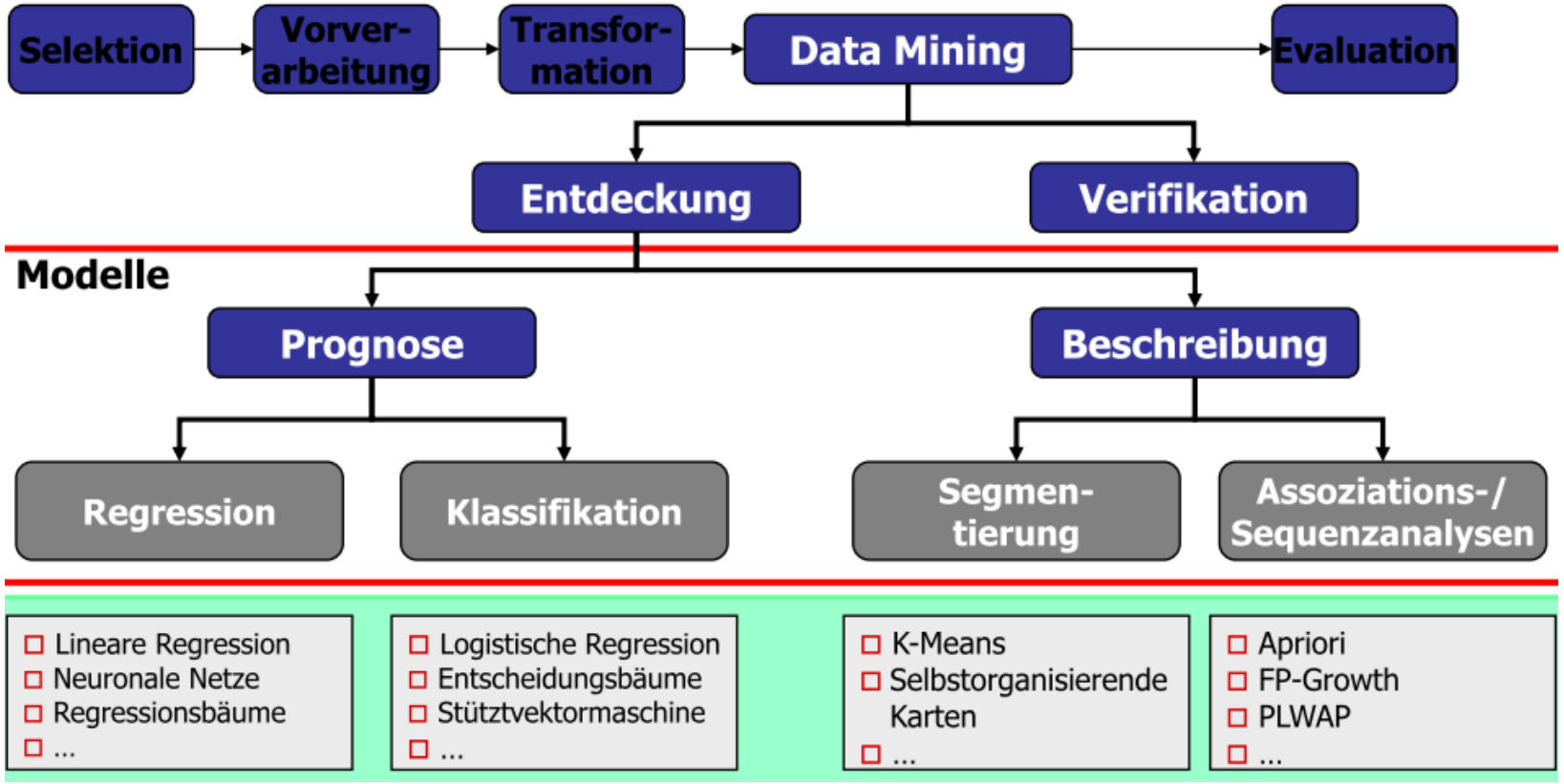
Auxiliary AI GmbH

Clustering with spaCy

- Block 05 -

Auxiliary AI GmbH

Machine Learning Übersicht



Auxiliary AI GmbH

Clustering

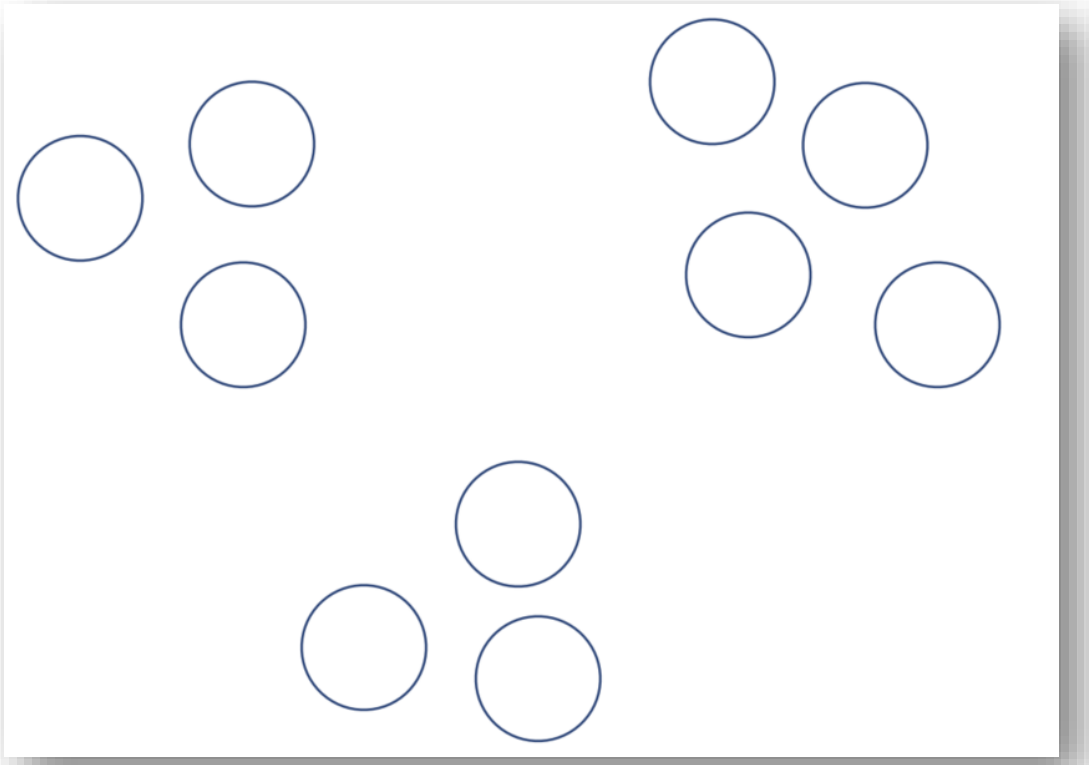
- Ihr habt Daten ... aber keine Label (y/Y)
- Was tun?

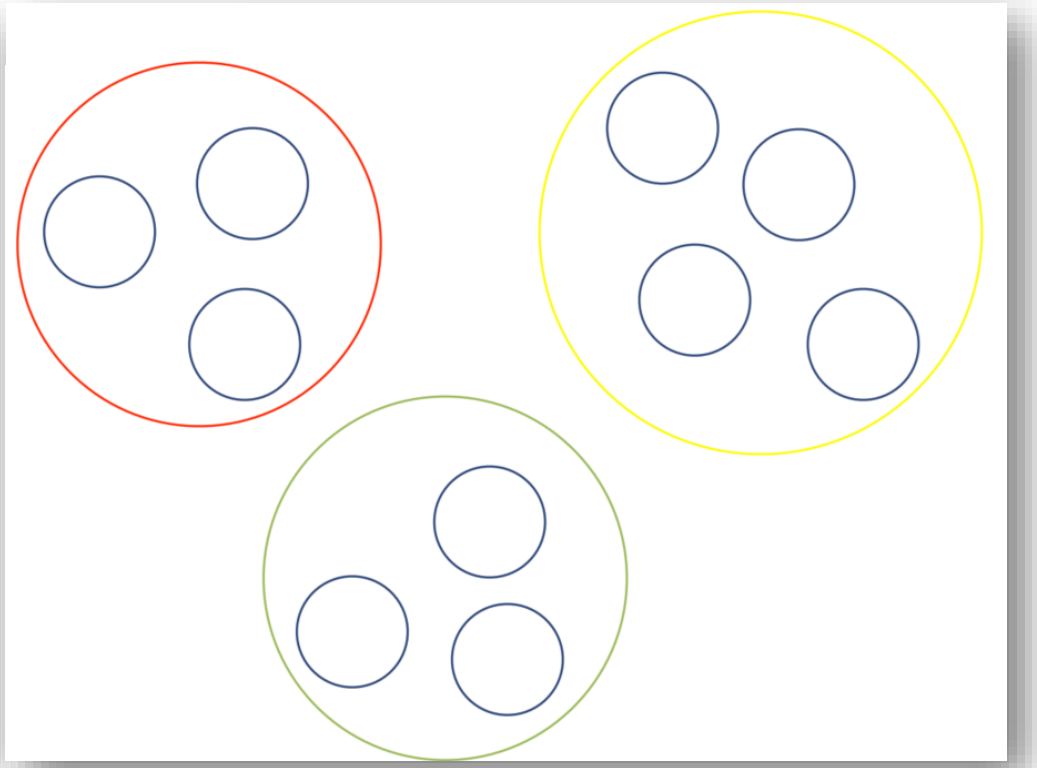
-> Clustering

- Nützlich, wenn nicht bekannt ist, wonach gesucht werden soll.

Auxiliary AI GmbH

Clustering

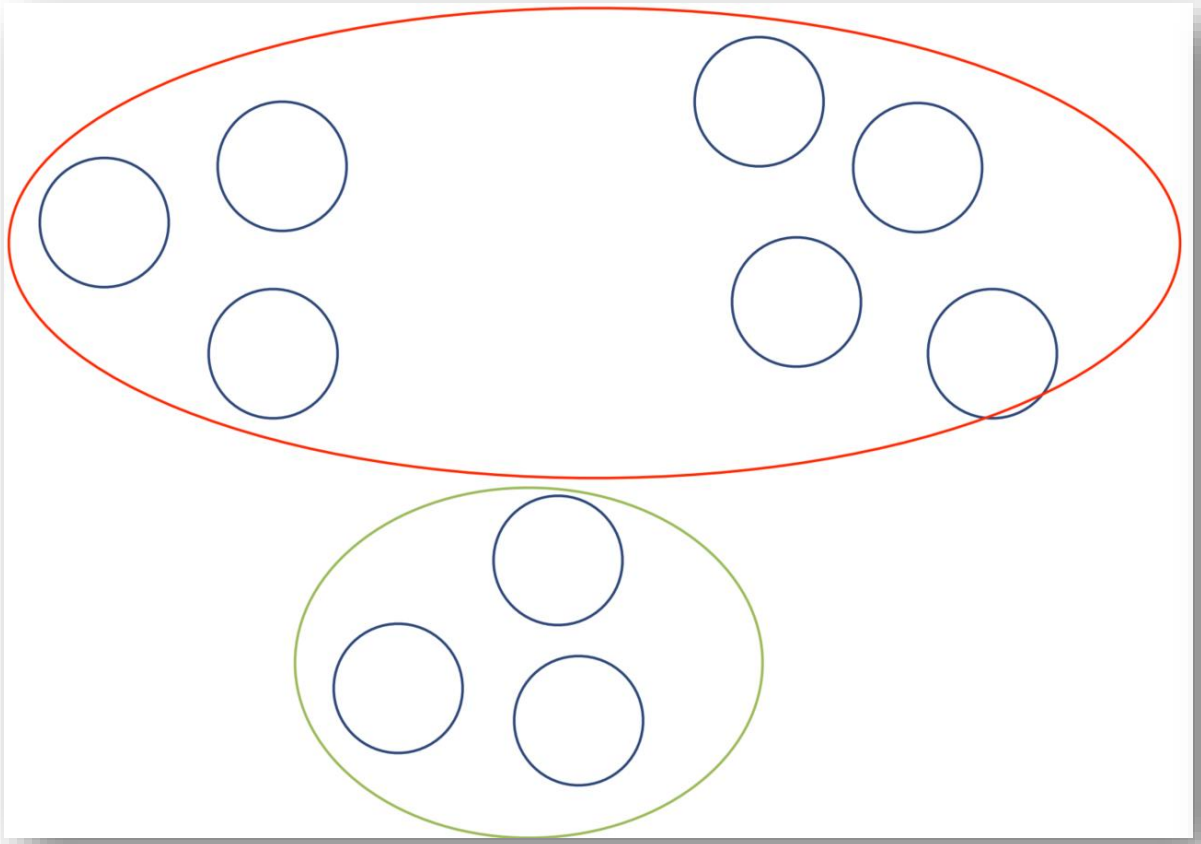


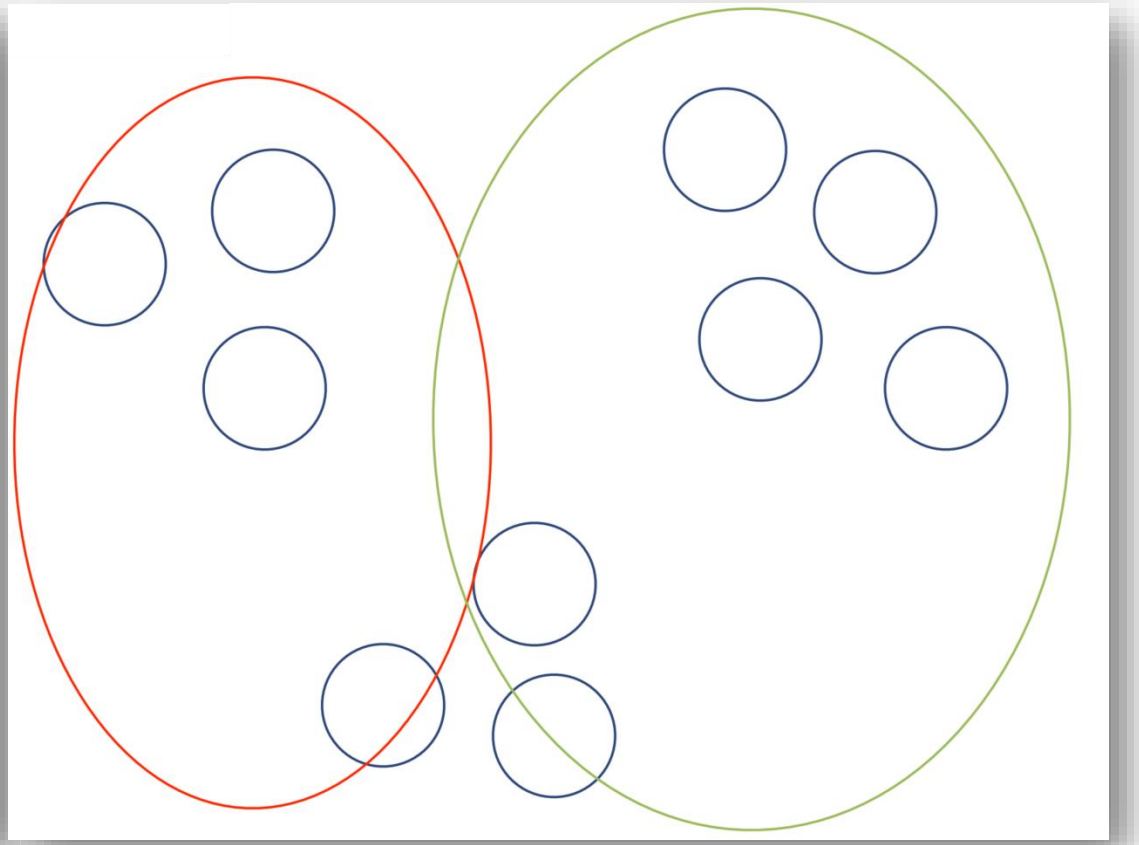


Picture from ojosujono96 (Freepik)

Auxiliary AI GmbH

Clustering

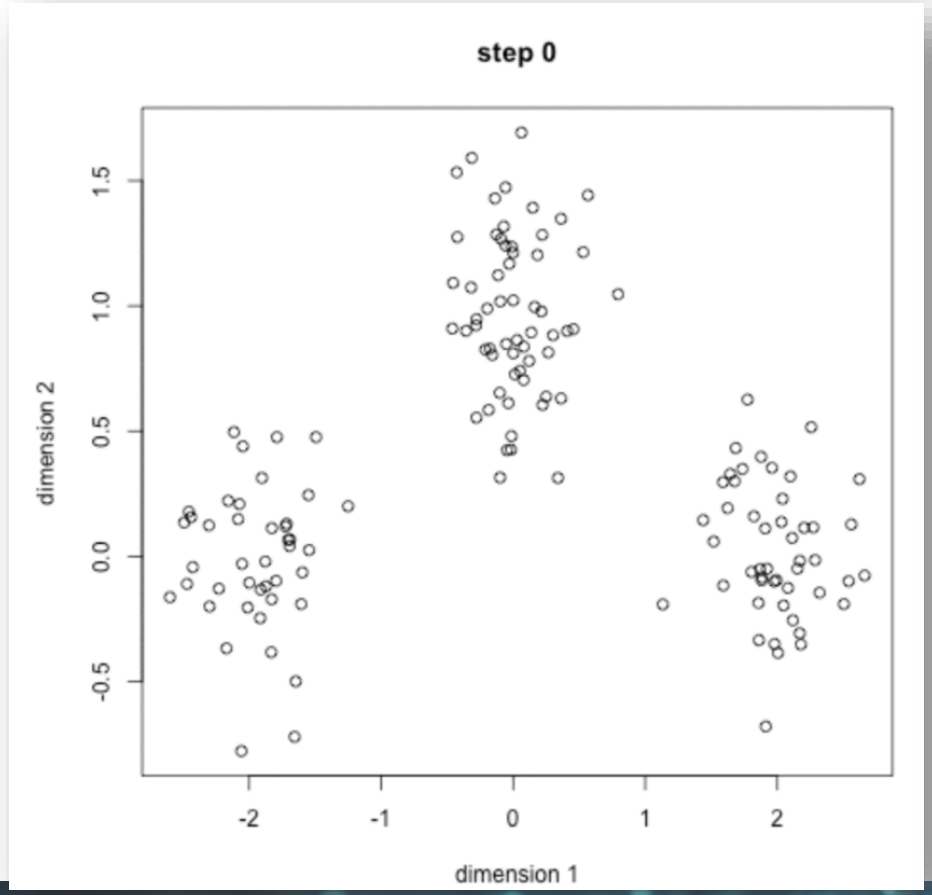


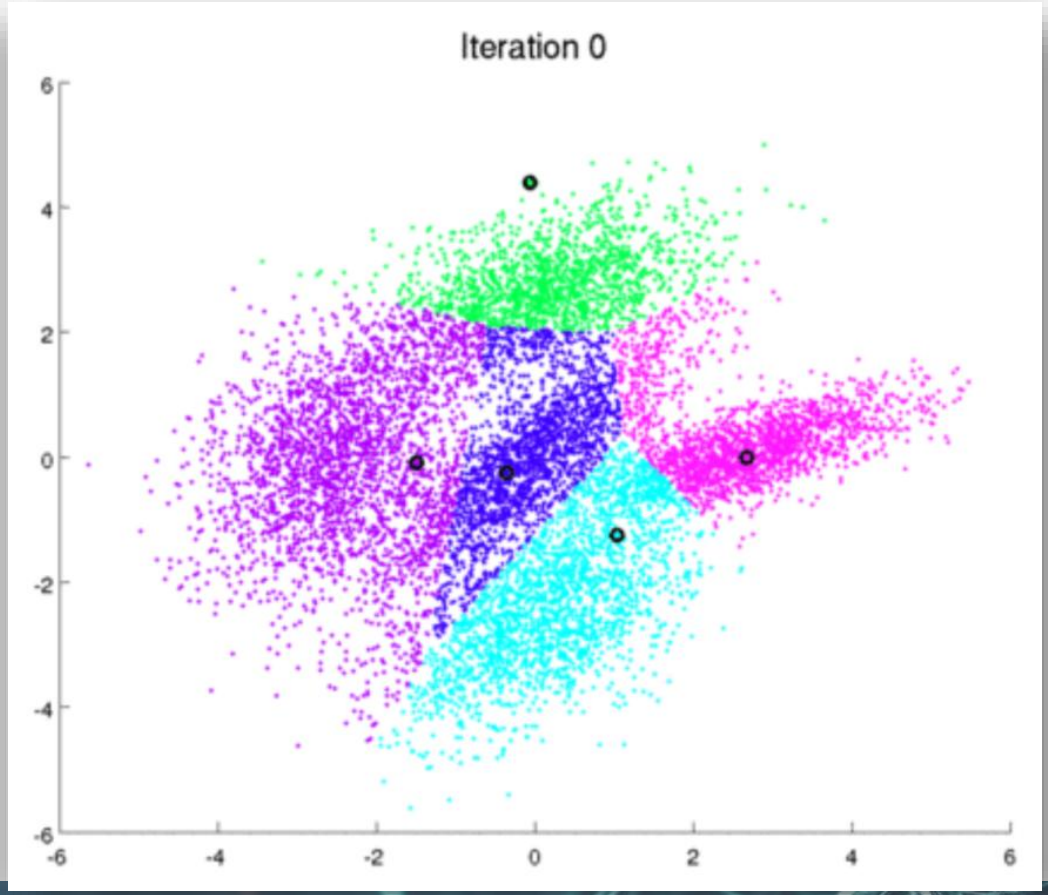


Picture from ojosujono96 (Freepik)

Auxiliary AI GmbH

Clustering





Picture from ojosujono96 (Freepik)

Auxiliary AI GmbH

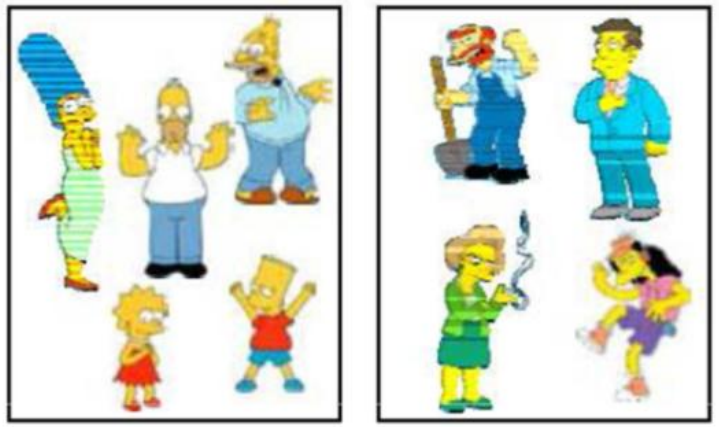
Clustering Considerations

- Hypothese:
  - Wir vermuten eine bestimmte Anzahl an Klustern in unseren Datensatz. Die Anzahl kann auf Erfahrungen aus der Praxis/Erhebung, Expertenmeinungen und/oder der Datenanalyse basieren.
- Vorgehen:
  - Wir wählen *k* Kluster und trainieren mehrere Modelle.
  - Anschließend vergleichen wir die Kluster über deren Ähnlichkeit und optimieren, damit diese eindeutiger sind. Sprich: Trennbar und damit wenig Überschneidungen aufweisen.

Clustering Types

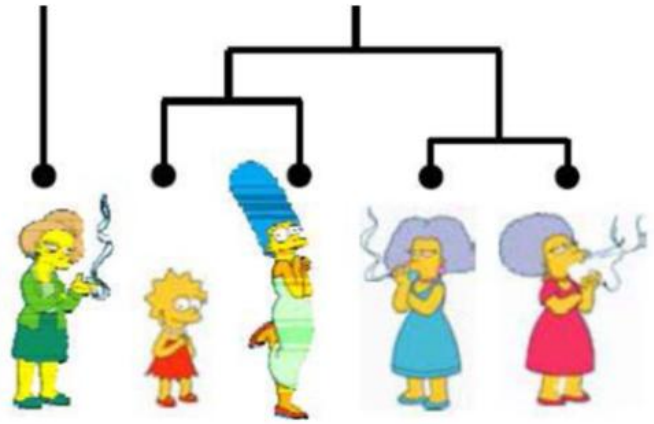
Partition algorithms (Flat)

- K-means
- Mixture of Gaussian
- Spectral Clustering



Hierarchical algorithms

- Bottom up – agglomerative
- Top down – divisive



Clustering for Images

- image segmentation
- break up the image into meaningful or perceptually similar regions

Auxiliary AI GmbH

Clustering Algorithms

- **KMeans**
  - Ein iterativer Algorithmus, der Daten in K Cluster aufteilt, indem er die Datenpunkte den nächstgelegenen Clusterzentroiden zuordnet. K-Means minimiert die Inertia.
- **DBScan** (Density-Based Spatial Clustering of Applications with Noise)
  - Ein dichtenbasierter Algorithmus, der Cluster basierend auf der Dichte der Datenpunkte identifiziert und Rauschen (Ausreißer) als nicht zugehörig betrachtet.

Clustering Inertia

- **Definition**: Die Inertia ist ein Maß für die Summe der quadrierten Abstände zwischen den Datenpunkten und den Zentroiden ihrer jeweiligen Cluster. Sie gibt also an, wie nah die Punkte innerhalb eines Clusters zueinander sind.

$$I = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - \mu_i||^2$$

- Hier ist k die Anzahl der Cluster, C_i ist das i-te Cluster, x_j sind die Datenpunkte und μ_i ist der Zentroid des i-ten Clusters.

Auxiliary AI GmbH

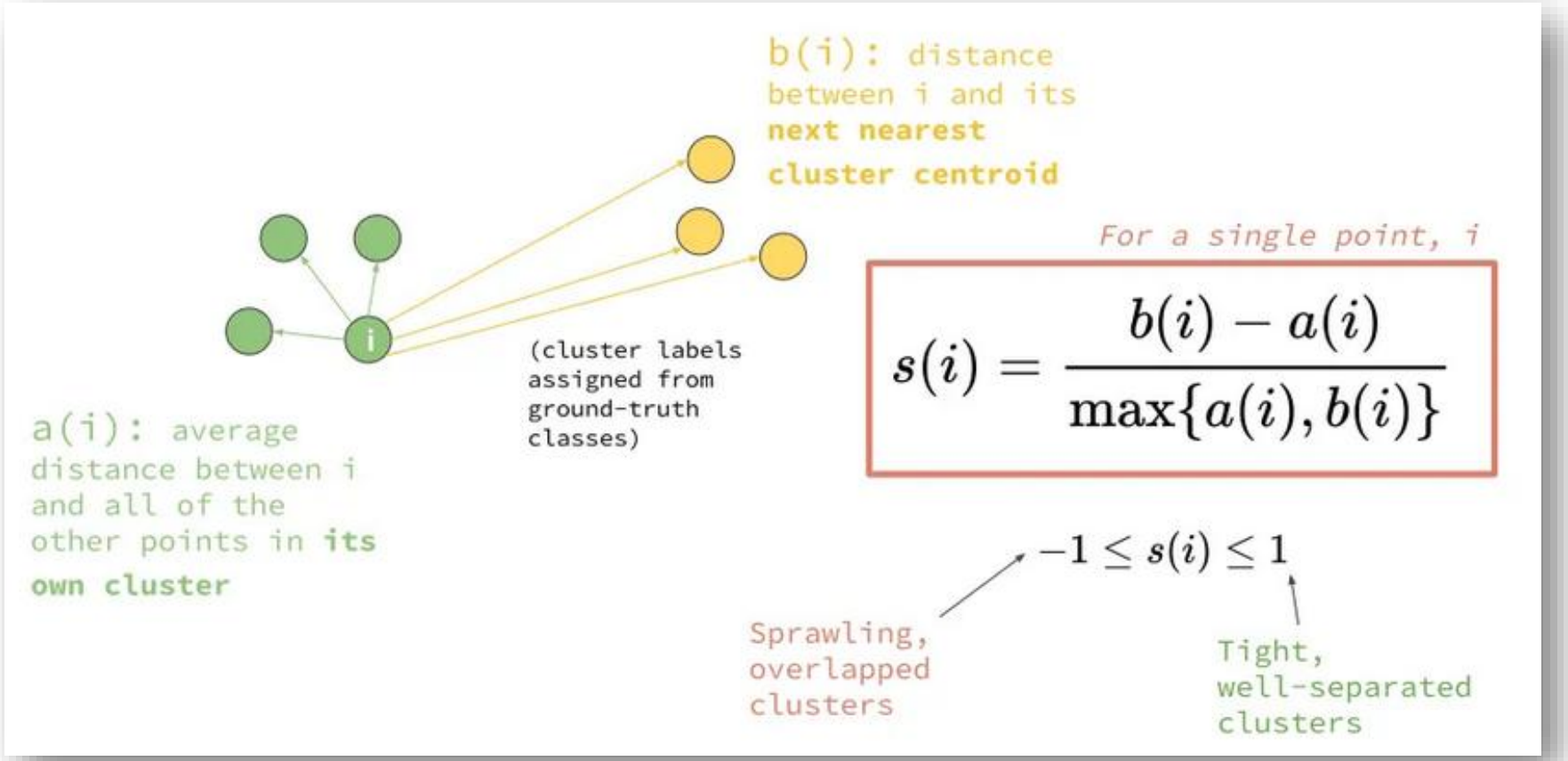
Clustering Silhouette Score

- **Silhouette Score**:
  - Misst die Qualität der Clusterung. Ein Wert zwischen -1 und 1 (je näher bei 1, desto besser). Er gibt an, wie gut ein Punkt zu seinem eigenen Cluster (Kohäsion) und wie gut er zu benachbarten Clustern (Trennung) passt.
  - Der Silhouette-Score für einen einzelnen Datenpunkt misst, wie gut der Punkt in seinem Cluster im Vergleich zu anderen Clustern "platziert" ist und kann als eine Art "Vertrauensmaß" für die Clusterzugehörigkeit angesehen werden.

1. **Cohesion (a(i))**: How close a data point is to other points in its own cluster.
2. **Separation (b(i))**: How far a data point is from the points in the nearest neighboring cluster.

Auxiliary AI GmbH

Clustering Silhouette Score



Picture from ojosujono96 (Freepik)

Auxiliary AI GmbH

Clustering Davies-Boludin Score/Index

- Der Davies-Bouldin-Index (DB-Index) ist ein Validierungsmaß, das häufig zur Evaluierung von Clustering-Ergebnissen verwendet wird. Er misst die Qualität eines Clusters, indem er das Verhältnis der internen Cluster-Distanz zur maximalen Distanz zwischen Clustern betrachtet. Ein niedrigerer Wert des Davies-Bouldin-Index deutet auf eine bessere Separation zwischen den Clustern und eine höhere Homogenität innerhalb der Cluster hin.

Auxiliary AI GmbH

Clustering Davies-Boludin Score/Index

Für einen Datensatz $X = X_1, X_2, X_3, ...$ kann der Davies-Bouldin-Index für k Cluster wie folgt berechnet werden:

$$D\ B = \frac{1}{k}\Sigma_{ich\ =\ 1}^{k} max\ x\ (\frac{\Delta\ (\ X_{ich}) + \Delta\ (\ X_J)}{\delta\ (\ X_{ich},\ X_J)})$$

Wo:

$\Delta\ X_k$ ist die Intracluster-Distanz innerhalb des Clusters X_k.

$\delta\ (\ X_{ich}, X_J)$ ist die Intercluster-Distanz zwischen den Clustern $X_{ich}\ Und\ X_J$.

- https://www.geeksforgeeks.org/davies-bouldin-index/

Auxiliary AI GmbH

Auxiliary AI GmbH

Geschäftsführer	**Marten Borchers & Benjamin Klinkigt**
Anschrift	Am Ziegelteich 74 22525 Hamburg Deutschland
Handelsregister	HR B 185519
Registergericht	Amtsgericht Hamburg
Umsatzsteuer-ID	DE 366 814 276
Kontakt	info@auxiliary-ai.de
Webseite	https://auxiliary-ai.de/