Bitte beachten Sie die Lizenzhinweise auf der letzten Seite

Foto: Omar Flores auf Unsplash

# Clusteranalyse – Evaluierung

## Teil 1: Intrinsische Evaluierung

INNOVET
KI B³
Künstliche Intelligenz in die berufliche Bildung bringen

GEFÖRDERT VOM
INNOVET
Bundesministerium für Bildung und Forschung
bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

# Bewertung von Ergebnissen beim Clustern

Verschiedene Initialisierungen können unterschiedliche Ergebnisse liefern.
Um ein „bestes“ Ergebnis auszuwählen, braucht man ein Evaluationsmaß.

- Clusteranalyse gehört zu den nicht überwachten Lernverfahren
- D.h. es gibt kein „korrektes“ Ergebnis, mit dem man das Ergebnis vergleichen kann

„Intrinsische Evaluierung“

d.h. es werden keine äußeren Faktoren berücksichtigt

Bewerte, wie sauber die Cluster voneinander getrennt sind

Sind alle Punkte im passendsten Cluster?

# Intrinsische Evaluierung

# Variante 1: ein „gute“ Cluster-Struktur

- Idee:
  - Abstand zum selben Cluster: klein
  - Abstand zum nächsten Cluster: groß

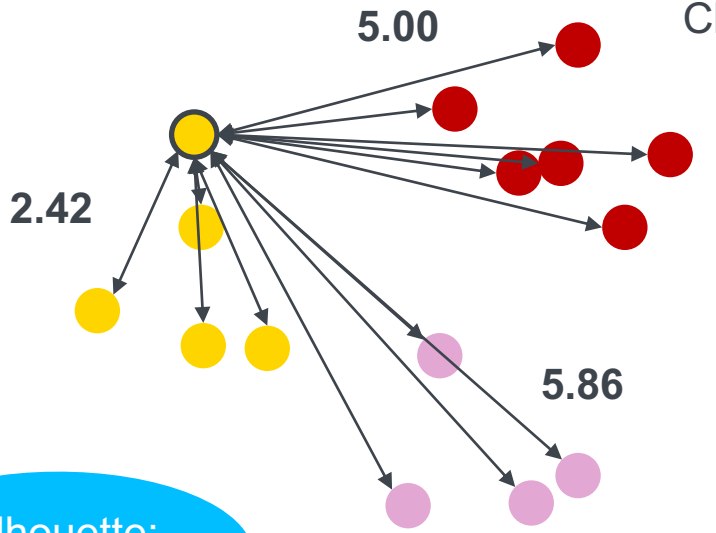
Abstand zum nächsten Cluster

Abstand zum selben Cluster

$$\frac{5.00 - 2.42}{5.00} \approx 0.52$$

geteilt durch den größeren Abstand

Silhouette: 0.52




Nächstes Cluster: rot

**Die Silhouette (auch: Silhouettenkoeffizient) eines Datenpunkts erfasst, wie gut er in sein Cluster passt.**

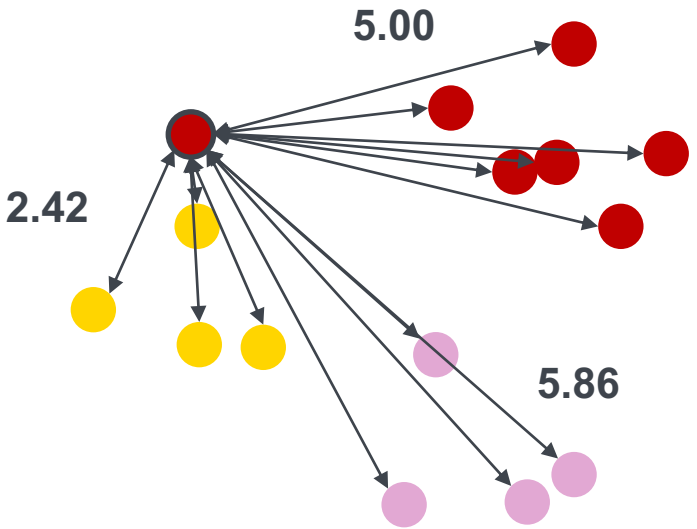
**Ein Wert von 1 drückt aus, dass der Datenpunkt perfekt in sein Cluster passt.**

**Man berechnet die Silhouette, indem man den mittleren Abstand zu Punkten des eigenen Clusters vom mittleren Abstand zu Punkten des nächsten Clusters abzieht und durch den größeren der beiden Werte teilt.**

# Variante 2: Ein „schlechte“ Cluster-Struktur

- Die Silhouette ist abhängig von der Zuordnung zum Cluster

$$\frac{2.42 - 5.00}{5.00} \approx -0.52$$

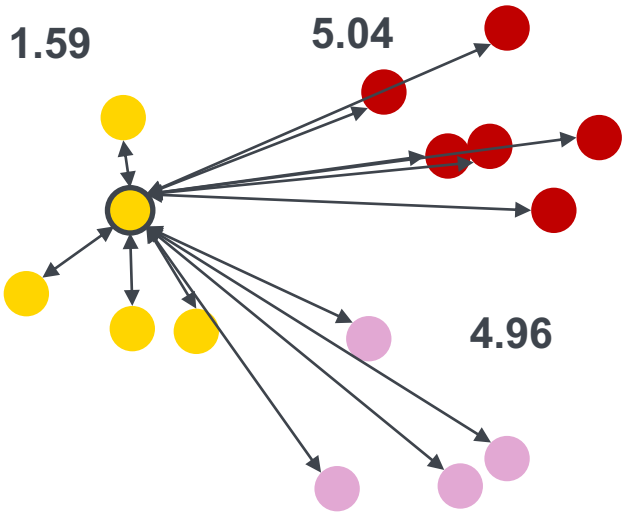


**Ein negativer Wert für die Silhouette bedeutet, dass der Punkt näher zu Punkten eines anderen Clusters liegt als zu den Punkten des eigenen Clusters.**

# Wann ist ein Punkt im „richtigen“ Cluster?

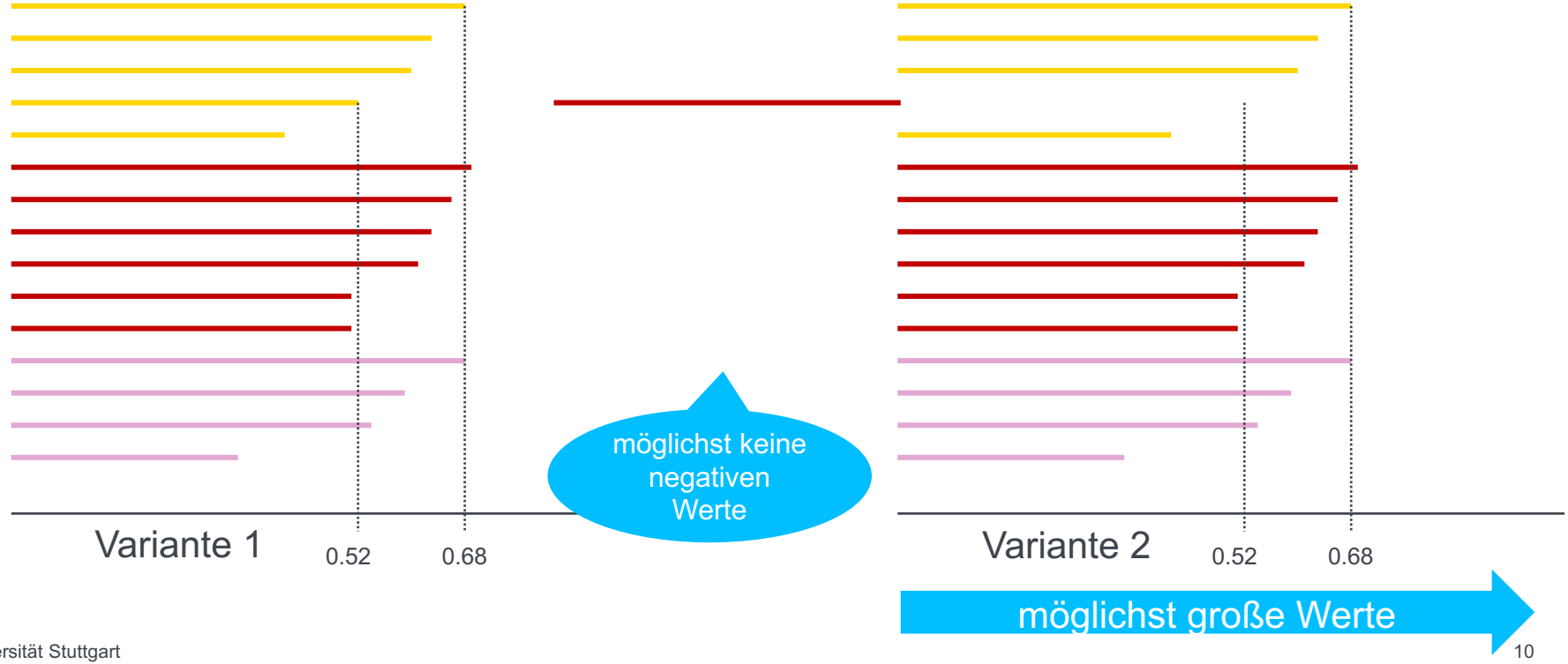
- Silhouette für einen Punkt mitten im Cluster

$$\frac{4.96 - 1.59}{4.96} \approx 0.68$$



# Der Silhouettenplot

Alle Silhouetten im Überblick

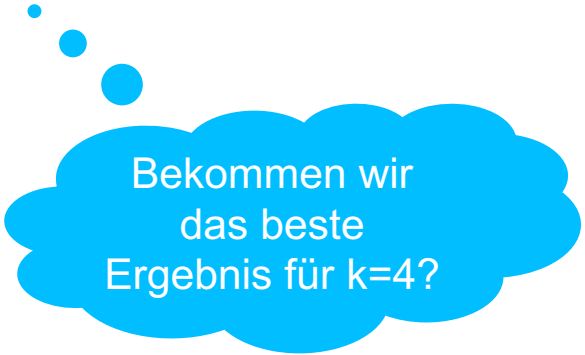


Der Silhouettenplot zeigt die Silhouetten aller Datenpunkte an, sortiert nach Clustern. Innerhalb der Cluster werden die Werte nach der Größe sortiert.

Clusterergebnisse sind gut, wenn alle Werte möglichst groß sind und (fast) keine negativen Werte vorkommen.

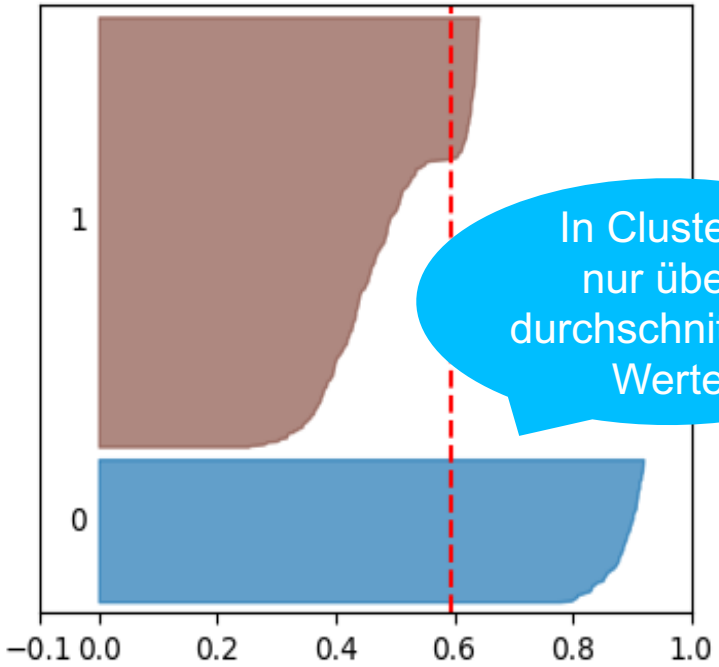
# Ein Experiment

- 4 Punktwolken zufällig erzeugen lassen
- Mit k-Means mit verschiedenen Werten für k clustern lassen
- Silhouetten-Plots und Mittelwerte vergleichen

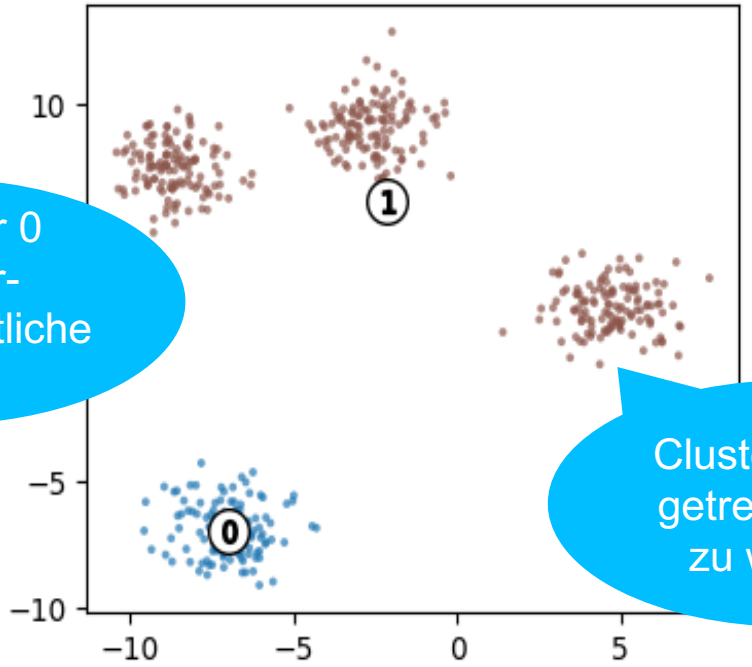


# Experiment 1, k=2

Keine negativen Werte, Mittelwert 0.60, Silhouette von Cluster 1 schlechter




In Cluster 0 nur über-durchschnittliche Werte



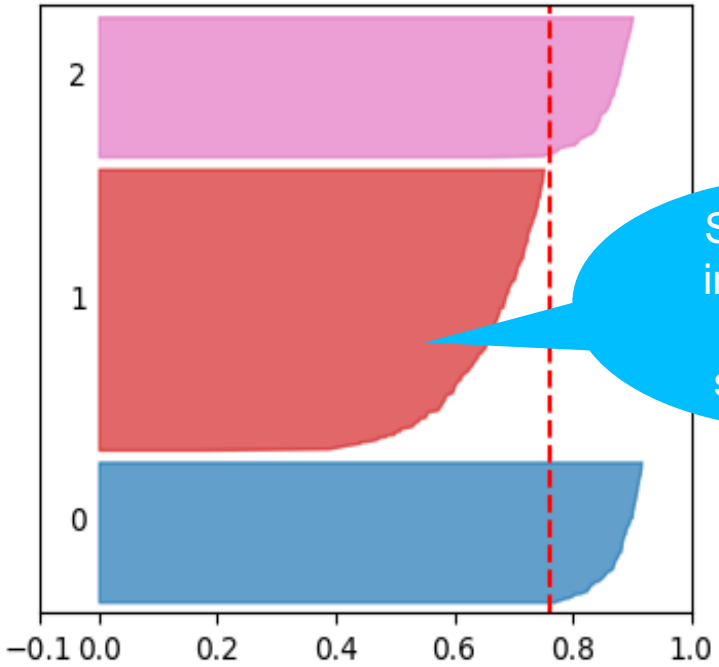

Cluster sauber getrennt (aber zu wenige)



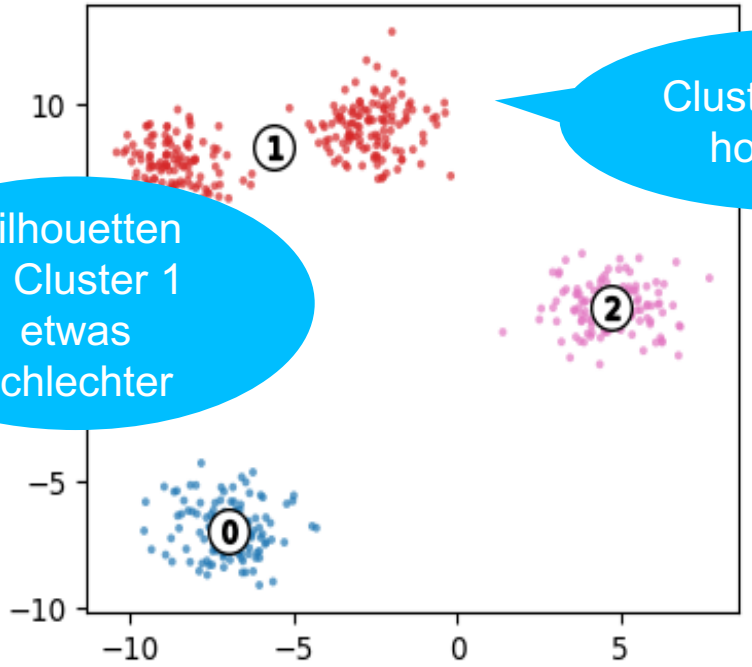
Bild generiert mit Code von https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

# Experiment 1, k=3

Mittelwert größer: 0.76




Silhouetten in Cluster 1 etwas schlechter




Cluster 1 nicht homogen

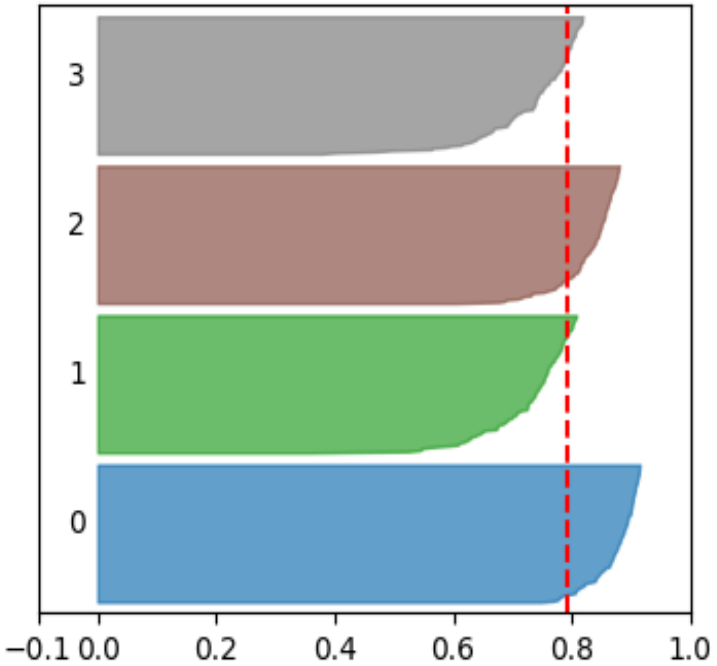


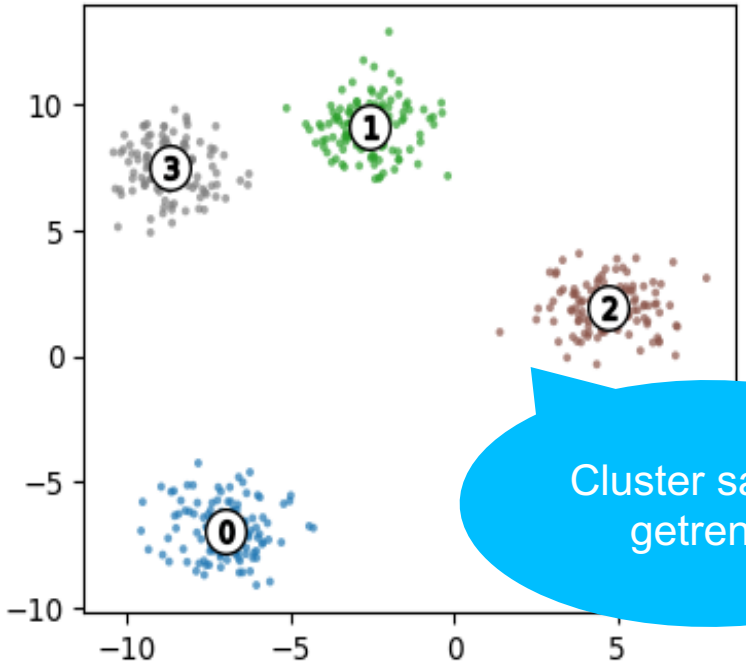
Bild generiert mit Code von https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

# Experiment 1, k=4

In allen Clustern kommen ähnlich gute Werte vor

Mittelwert noch größer: 0.79



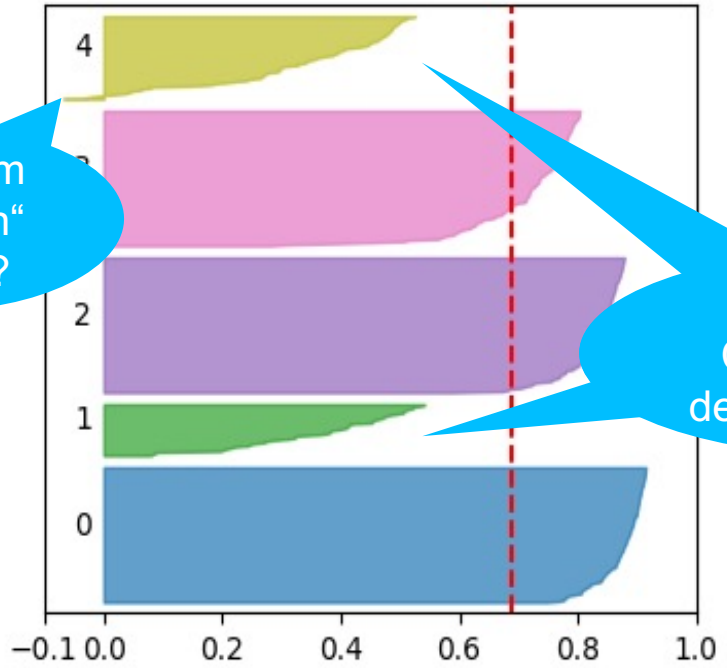

Cluster sauber getrennt



Bild generiert mit Code von https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

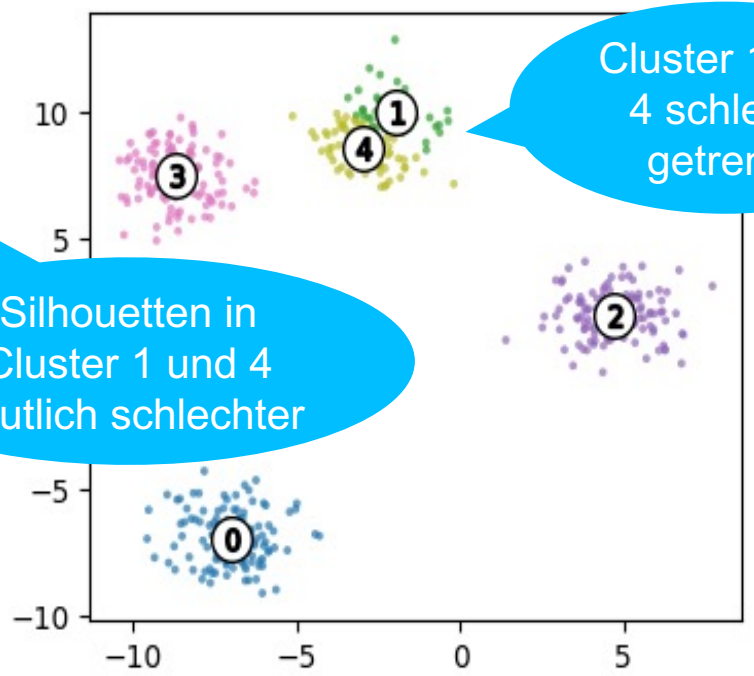
# Experiment 1, k=5

Silhouetten kleiner (Mittelwert: 0.69), einige negative Werte




Punkte im „falschen“ Cluster?

Silhouetten in Cluster 1 und 4 deutlich schlechter



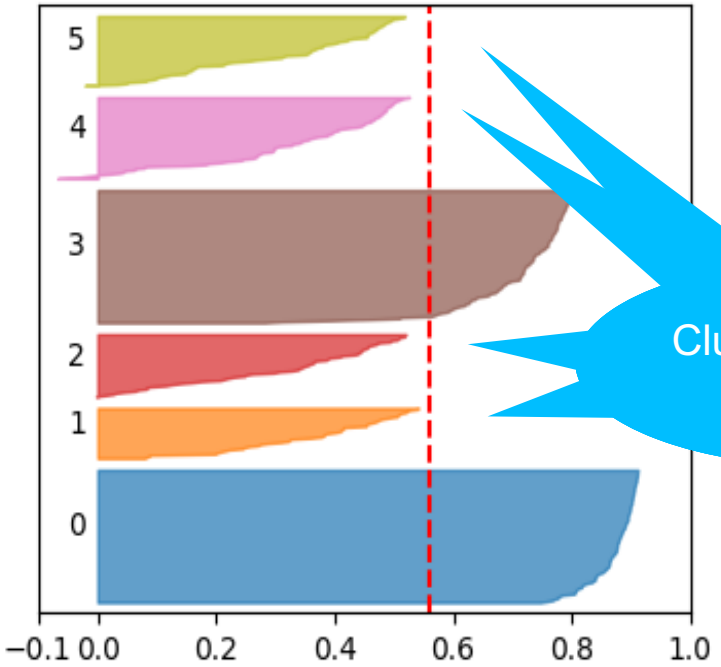

Cluster 1 und 4 schlecht getrennt

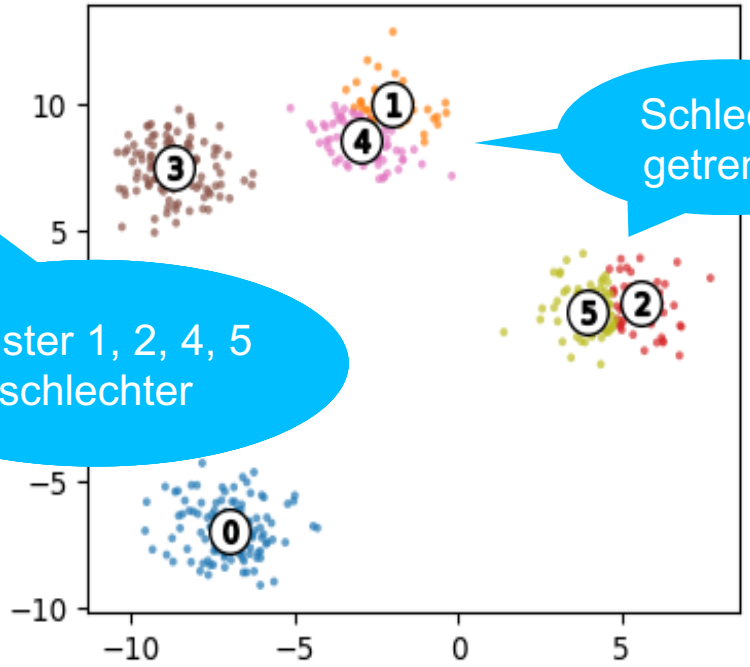


Bild generiert mit Code von https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

# Experiment 1, k=6

Silhouetten noch kleiner
(Mittelwert: 0.56)




Cluster 1, 2, 4, 5
schlechter

Schlecht
getrennt



Bild generiert mit Code von https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

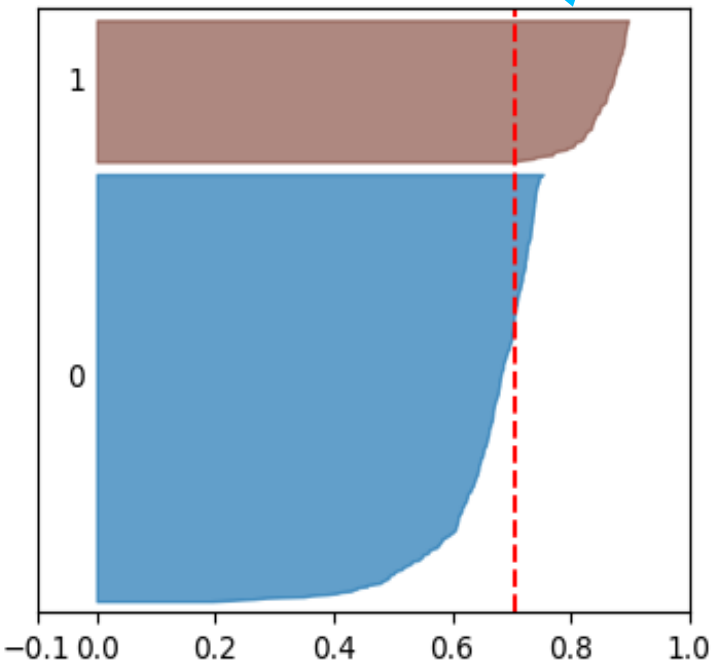
# Zusammenfassung Experiment 1

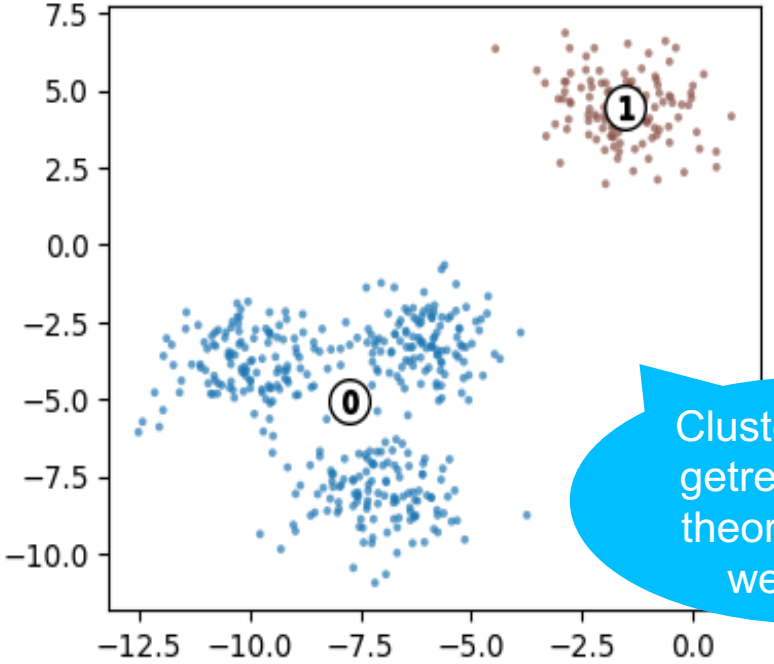
- Bei k=4
  - höchster Mittelwert
  - alle Cluster ähnlich „gut“
- Bei k=2 und k=3
  - mäßige Unterschiede zwischen den Silhouetten der Cluster
- Bei k=5 und k=6
  - negative Werte, deutlich niedrigere Mittelwerte
  - große Unterschiede zwischen den Silhouetten der Cluster

# Experiment 2, k=2

In Cluster 1 fast nur über-durchschnittliche Werte

Keine negativen Werte, Silhouetten relativ groß (Mittelwert: 0.70)



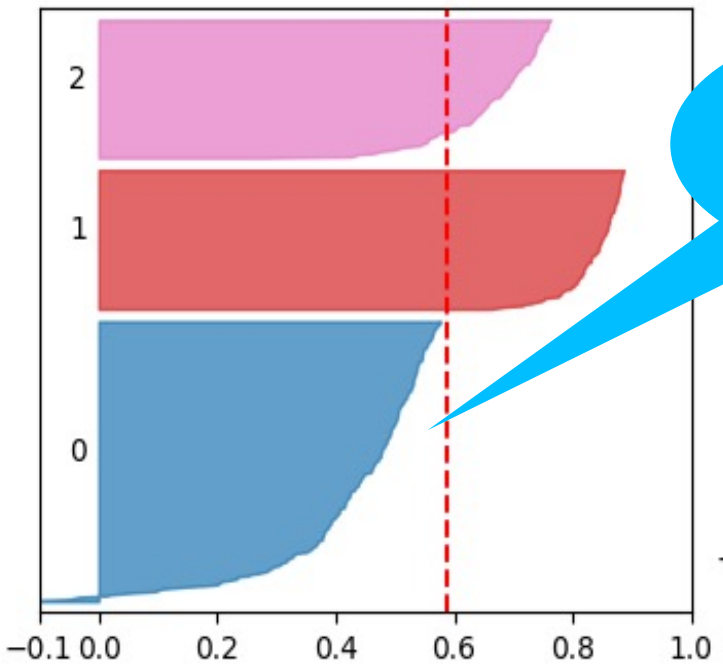

Cluster sauber getrennt (aber theoretisch zu wenige!!)



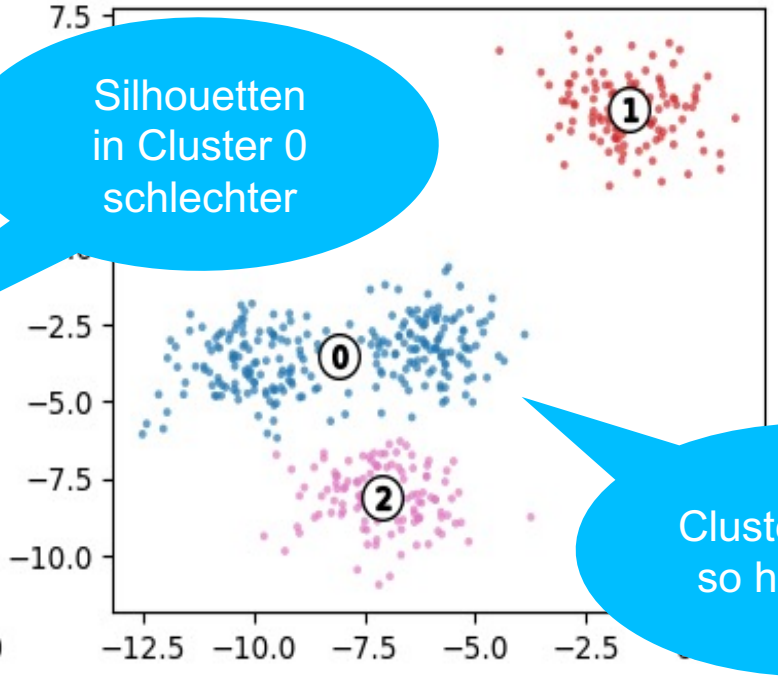
Bild generiert mit Code von https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

# Experiment 2, k=3

Negative Werte, Silhouetten insgesamt kleiner (Mittelwert: 0.59)




Silhouetten in Cluster 0 schlechter




Cluster 0 nicht so homogen

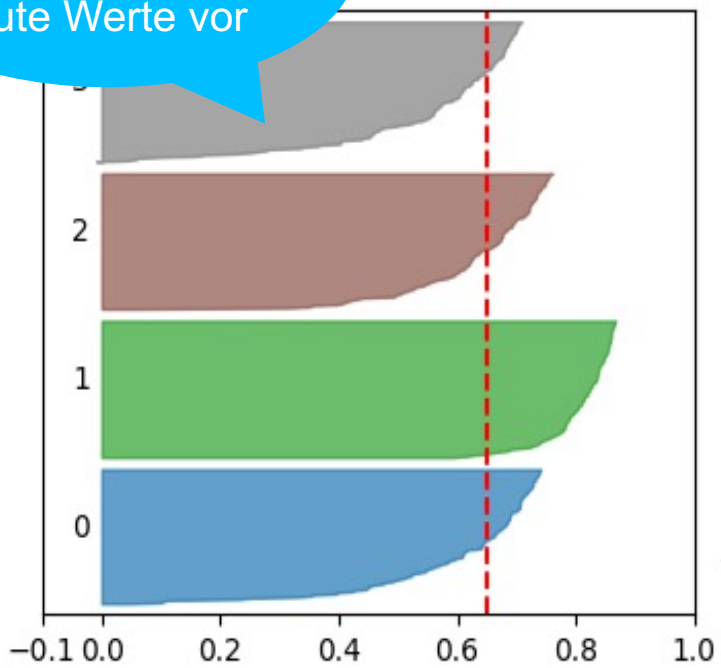


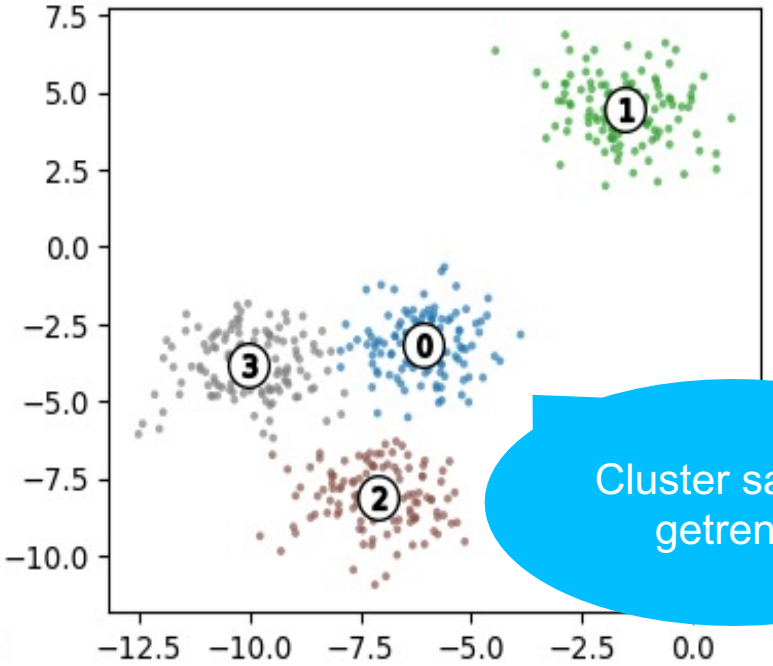
Bild generiert mit Code von https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

# Exp

In allen Clustern kommen ähnlich gute Werte vor

Keine negativen Werte, Silhouetten relativ groß (Mittelwert: 0.65)



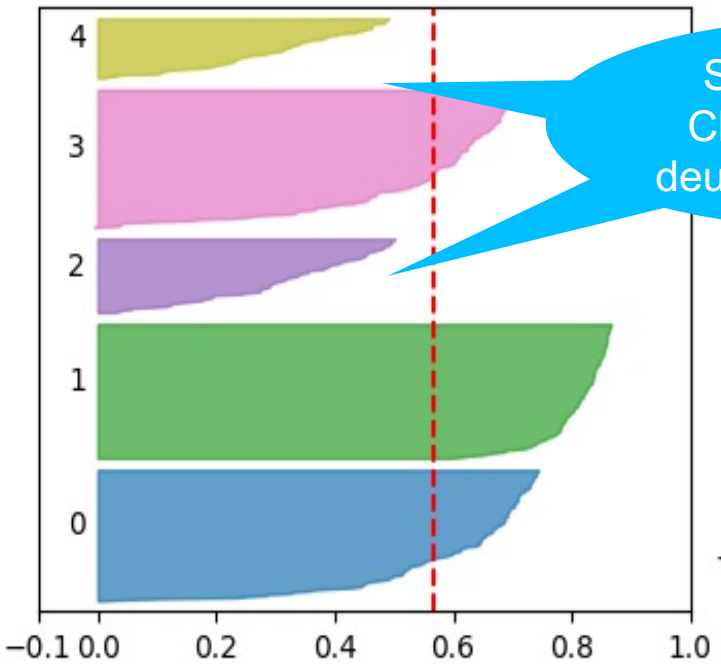

Cluster sauber getrennt



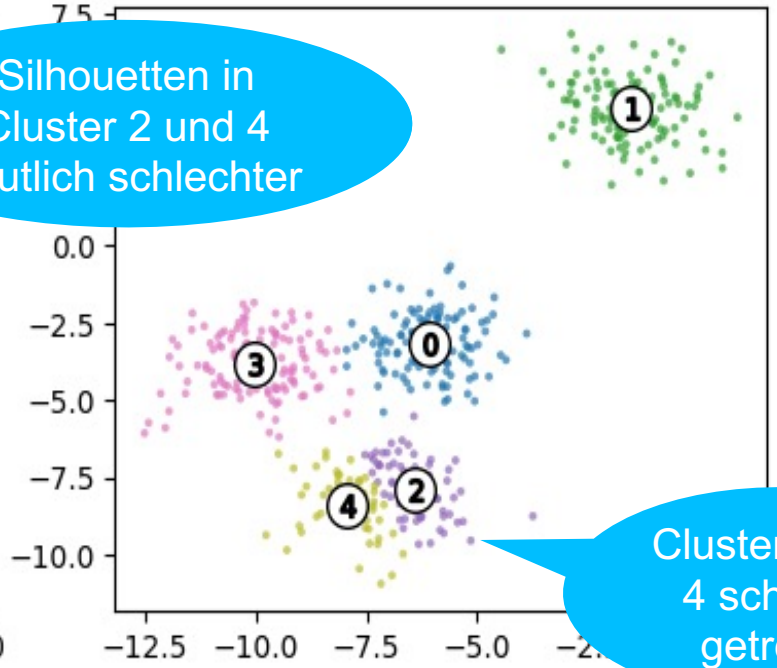
Bild generiert mit Code von https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

# Experiment 2, k=5

Keine negativen Werte, Silhouetten kleiner (Mittelwert: 0.56)




Silhouetten in Cluster 2 und 4 deutlich schlechter




Cluster 2 und 4 schlecht getrennt



Bild generiert mit Code von https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

# Experiment 2, k=6

Silhouetten noch kleiner
(Mittelwert: 0.48)

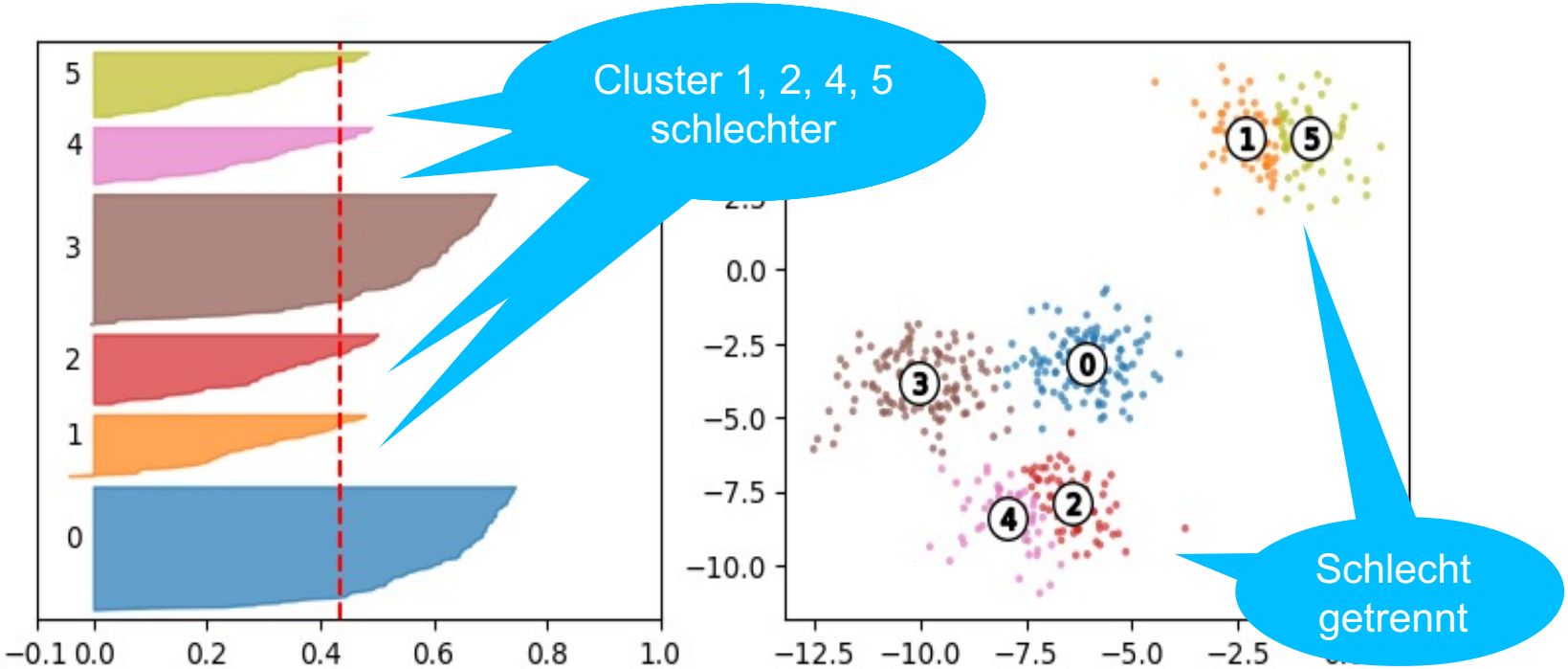





Bild generiert mit Code von https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

# Zusammenfassung Experiment 2

- Unerwartet: bei k=2 höchster Mittelwert!!!
  - vielleicht sinnvoll, da drei der vier Cluster nahe beieinander
- Immerhin: sowohl bei k=2 als auch bei k=4 keine negativen Werte
- Bei k=4 ähnlichste Silhouetten für alle Cluster

**Der Silhouetten-Mittelwert sowie Silhouettenplots dienen der intrinsischen Evaluierung von Clusterergebnissen.**

**Dabei sollte ein insgesamt hoher Mittelwert angestrebt werden. Außerdem ist es sinnvoll, Cluster zu vermeiden, die deutlich schlechtere Silhouetten aufweisen.**

# Silhouettenanalyse

- Kann zur intrinsischen Evaluierung verwendet werden
- Es gibt weitere Maße zu diesem Zweck!
- Silhouetten sind wahrscheinlich das bekannteste Maß

INNOVET

KI B³

Künstliche Intelligenz in die berufliche Bildung bringen


**Dr. Antje Schweitzer**

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung


**Universität Stuttgart**
Institut für Maschinelle Sprachverarbeitung
Institut für Software Engineering

IHK Industrie- und Handelskammer Reutlingen
Reutlingen | Tübingen | Zollernalb

IHK Region Stuttgart

IHK Industrie- und Handelskammer Karlsruhe

LMU Ludwig-Maximilians-Universität München


GEFÖRDERT VOM

INNOVET

Bundesministerium für Bildung und Forschung

bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

# Lizenzbestimmungen

“Clusteranalyse – Evaluierung, Teil 1: Intrinsische Evaluierung" von Antje Schweitzer, KI B³ / Uni Stuttgart

Das Werk - mit Ausnahme der folgenden Elemente:

- Logos der Verbundpartner und des Förderprogramms
- im Quellenverzeichnis aufgeführte Medien

ist lizenziert unter:

CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.de)
(Namensnennung 4.0 International)

**Quellenverzeichnis**

Titelfoto: Omar Flores (https://unsplash.com/de/@designedbyflores), ohne Titel, auf Unsplash (https://unsplash.com/de/fotos/blaue-rote-und-weisse-grafiken-lQT_bOWtysE), lizenziert unter Unsplash-Lizenz (https://unsplash.com/license). Bildausschnitt verändert.