Data Science 2

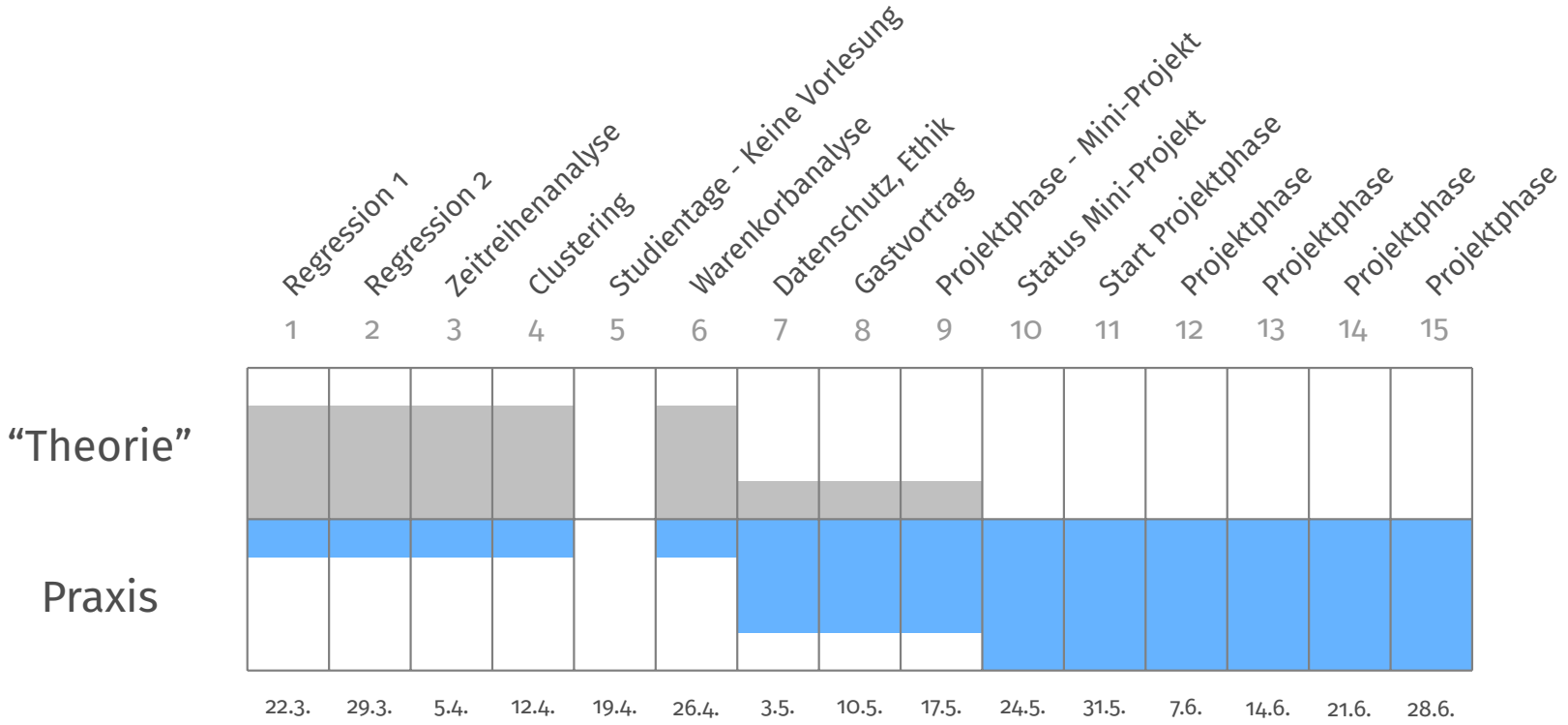
Vorlesung 4 - Intro

Prof. Dr. Christian Bockermann

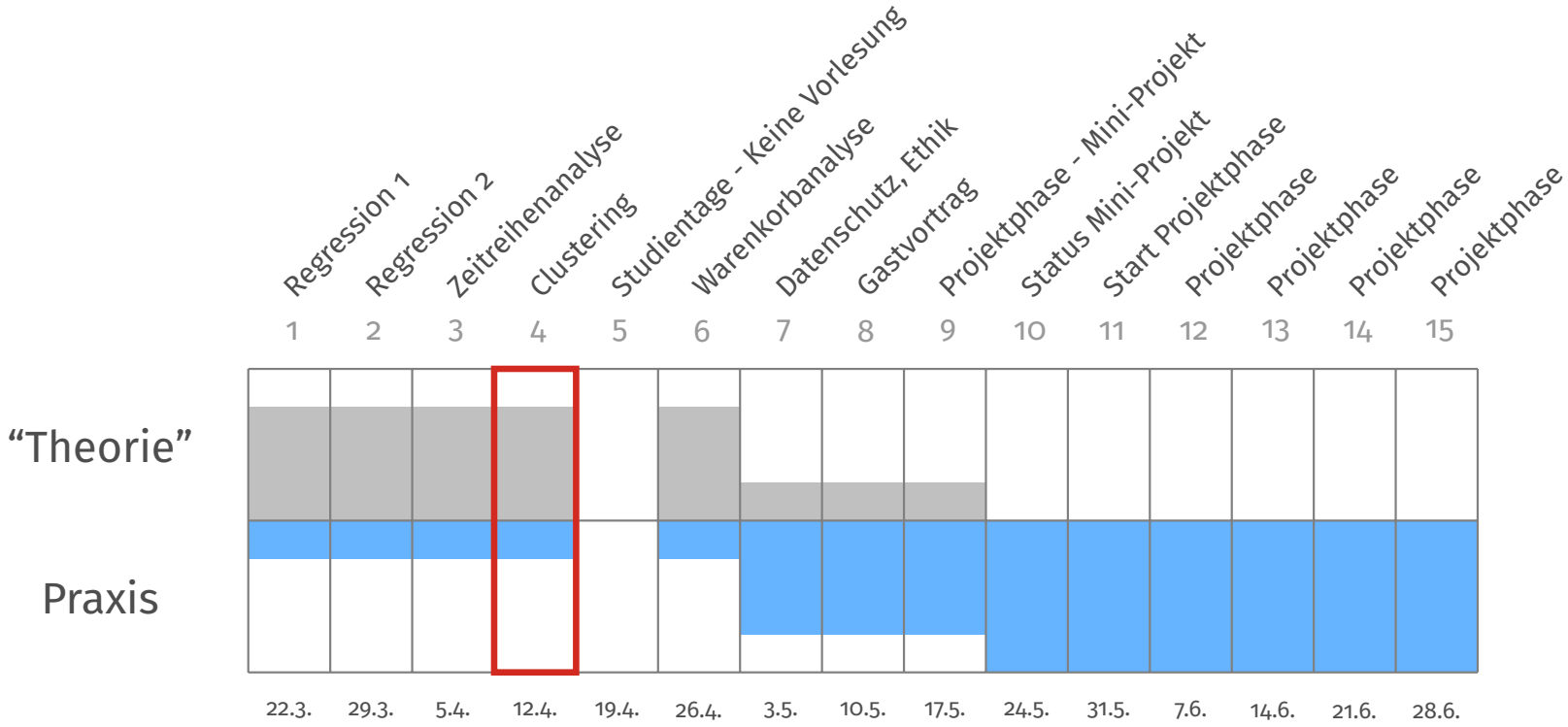
Hochschule Bochum

Sommersemester 2022

Themen der Vorlesung



Themen der Vorlesung



Wiederholung - Clustering

Bisher: Überwachtes Lernen mit Zielvariable

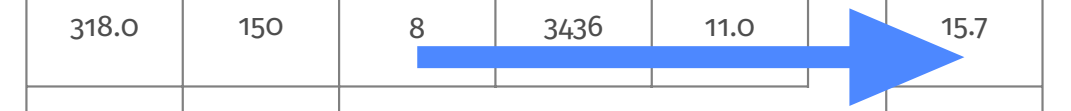
$X_{Hubraum}$	$X_{Leistung}$	$X_{Zylinder}$	$X_{Gewicht}$	$X_{Beschl.}$	$Y_{l/100km}$
307.0	130	8	3504	12.0	15.7
350.0	165	8	3693	11.5	18.8
318.0	150	8	3436	11.0	15.7
304.0	140	8	3433	12.0	17.6

Bisher: Überwachtes Lernen mit Zielvariable

$X_{Hubraum}$	$X_{Leistung}$	$X_{Zylinder}$	$X_{Gewicht}$	$X_{Beschl.}$	$Y_{l/100km}$
307.0	130	8	3504	12.0	15.7
350.0	165	8	3693	11.5	18.8
318.0	150	8	3436	11.0	15.7
304.0	140	8	3433	12.0	17.6

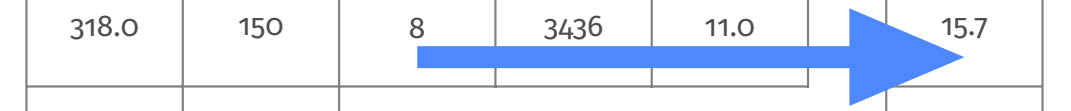
Bisher: Überwachtes Lernen mit Zielvariable

$X_{Hubraum}$	$X_{Leistung}$	$X_{Zylinder}$	$X_{Gewicht}$	$X_{Beschl.}$	$Y_{l/100km}$
307.0	130	8	3504	12.0	15.7
350.0	165	8	3693	11.5	18.8
318.0	150	8	3436	11.0	15.7
304.0	140	8			17.6



Bisher: Überwachtes Lernen mit Zielvariable

$X_{Hubraum}$	$X_{Leistung}$	$X_{Zylinder}$	$X_{Gewicht}$	$X_{Beschl.}$	$Y_{l/100km}$
307.0	130	8	3504	12.0	15.7
350.0	165	8	3693	11.5	18.8
318.0	150	8	3436	11.0	15.7
304.0	140	8			17.6



Bisher: Überwachtes Lernen mit Zielvariable

$X_{Hubraum}$	$X_{Leistung}$	$X_{Zylinder}$	$X_{Gewicht}$	$X_{Beschl.}$	$Y_{l/100km}$
307.0	130	8	3504	12.0	15.7
350.0	165	8	3693	11.5	8.
318.0	150	8	3436	11.0	1 7
304.0	140	8	3433	12.0	7.

Unüberwachtes Lernen: Es gibt keine Vorhersage-Variable!

Clustering sucht Aufteilung von Daten in ähnliche Gruppen

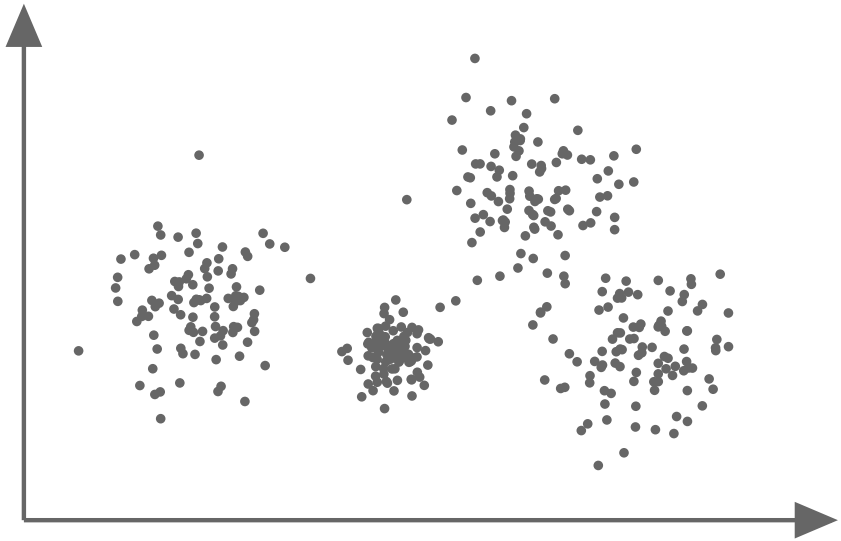
- Datenmenge **X** von Beispielen (keine Klassen gegeben!)
- Parameter k zu findender Gruppen
- Abstandsmaß $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- Qualitätsfunktion q

Ziel:

- Abstand *innerhalb* der Gruppen soll minimiert, Abstand *zwischen* den Gruppen soll maximiert werden

Beispiel: Clustering

Sei $\mathbf{C} = C_1, \ldots, C_k$ eine Aufteilung der Daten X (ein *Clustering*)

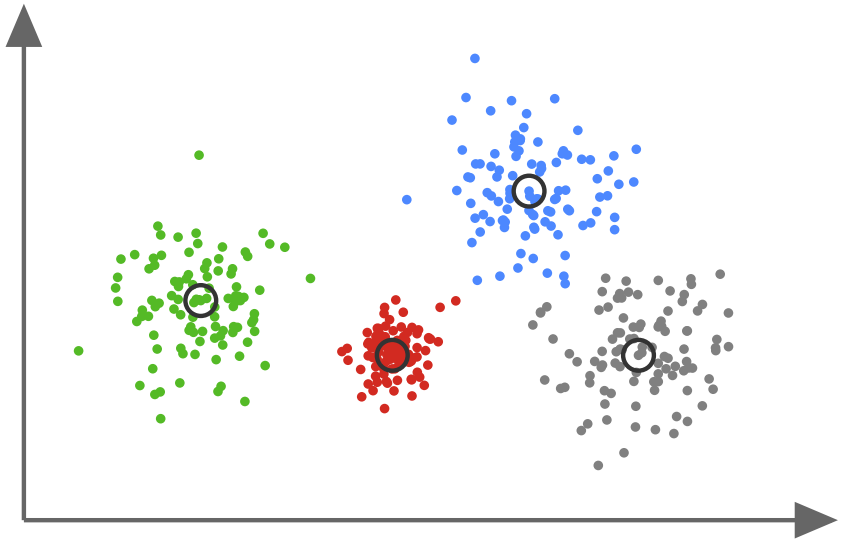


Qualitätsfunktion: (Innere Abstände)

$$q_{inner}(\mathbf{C}) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \bar{\mathbf{c}}_i) \quad \text{, mit } \bar{\mathbf{c}}_i \text{ Zentrum von } C_i$$

Beispiel: Clustering

Sei $\mathbf{C} = C_1, \ldots, C_k$ eine Aufteilung der Daten X (ein *Clustering*)



Clustering auf Datenpunkten mit $k = 4$. Die schwarzen Kreise markieren jeweils das Zentrum $\bar{\mathbf{c}}_i$ des jeweiligen Cluster C_i.

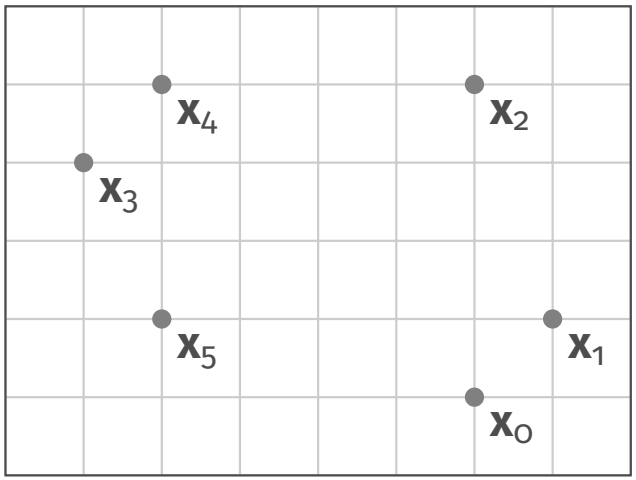
Qualitätsfunktion: (Innere Abstände)

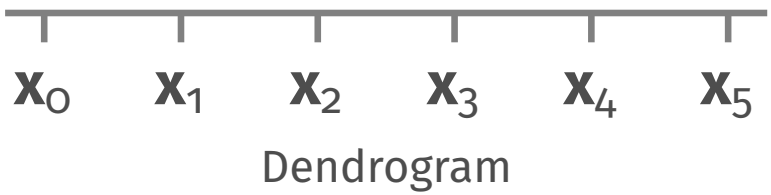
$$q_{inner}(\mathbf{C}) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \bar{\mathbf{c}}_i) \quad \text{, mit } \bar{\mathbf{c}}_i \text{ Zentrum von } C_i$$

Einordnung von Clustering-Verfahren

- Hierarchisches Clustering, agglomerativ/divisiv
- Iterative Verfahren (z.B. k-Means)
- Dichte-basiertes Clustering (z.B. DBScan)
- Stochastische Verfahren
- Meta-Daten Basierte Verfahren

Hierarchisches Clustering (agglomerativ)

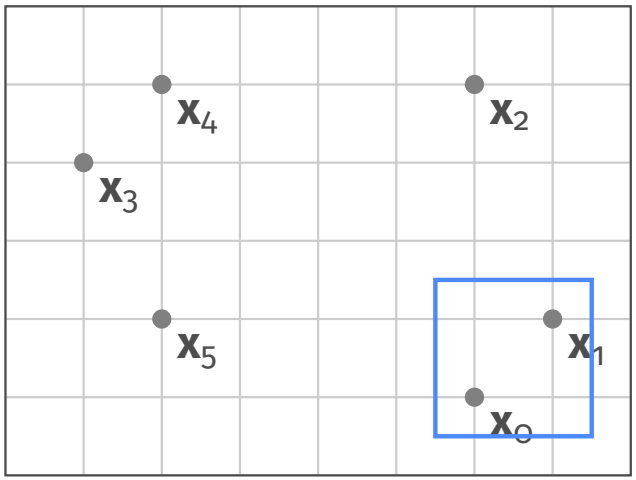


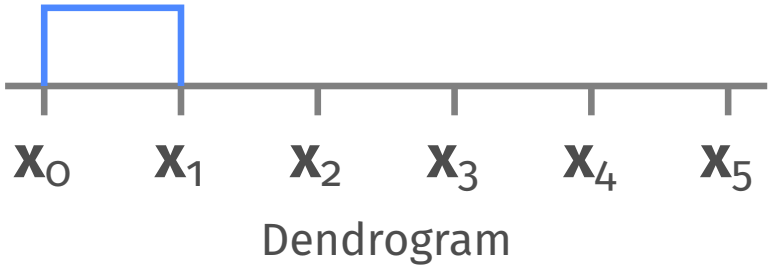


Dendrogram

- Jeder Datenpunkt ist anfangs ein Cluster
- In jedem Schritt werden die nächstgelegenen Cluster zusammengefasst
- Erzeugt Hierarchie von Aufteilungen der Daten

Hierarchisches Clustering (agglomerativ)

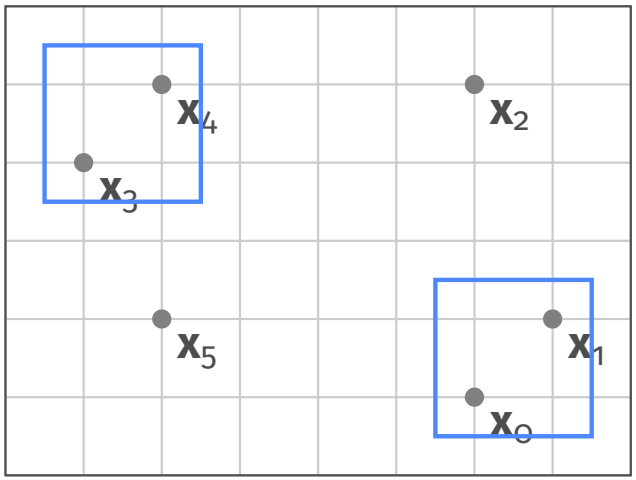


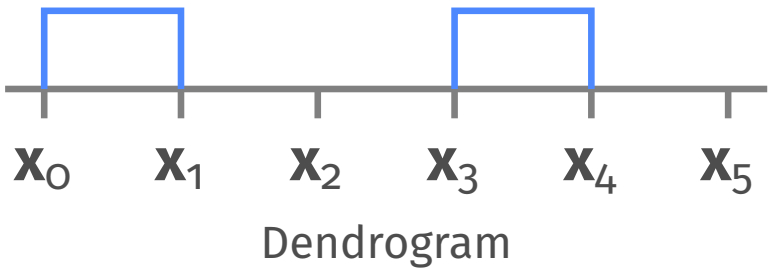


Dendrogram

- Jeder Datenpunkt ist anfangs ein Cluster
- In jedem Schritt werden die nächstgelegenen Cluster zusammengefasst
- Erzeugt Hierarchie von Aufteilungen der Daten

Hierarchisches Clustering (agglomerativ)

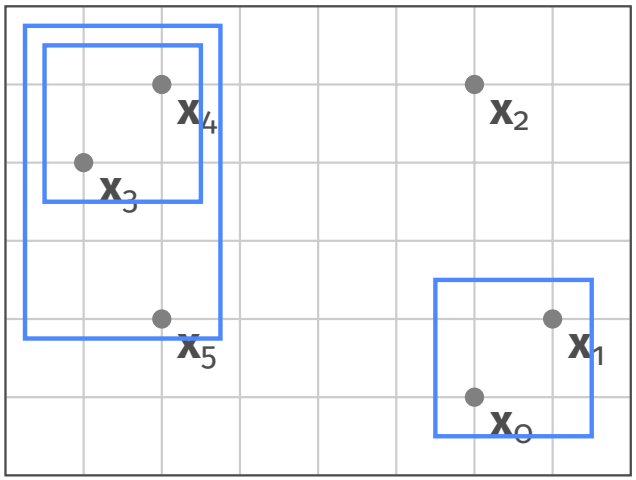


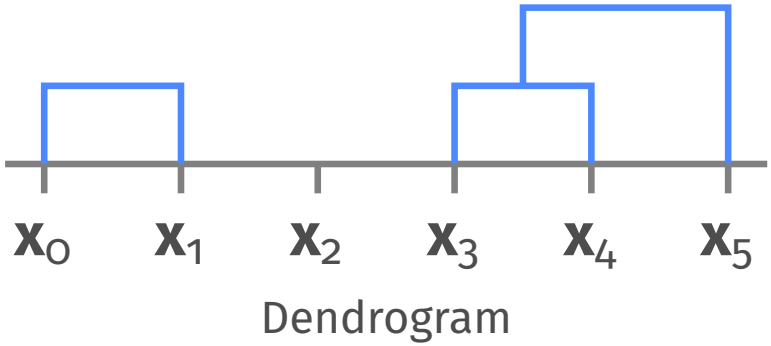


Dendrogram

- Jeder Datenpunkt ist anfangs ein Cluster
- In jedem Schritt werden die nächstgelegenen Cluster zusammengefasst
- Erzeugt Hierarchie von Aufteilungen der Daten

Hierarchisches Clustering (agglomerativ)

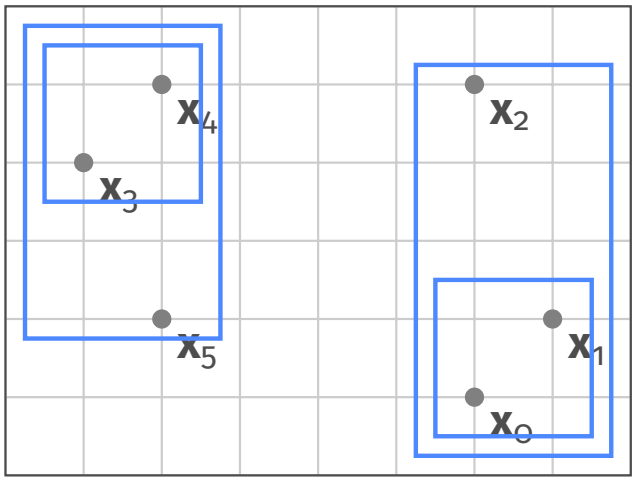


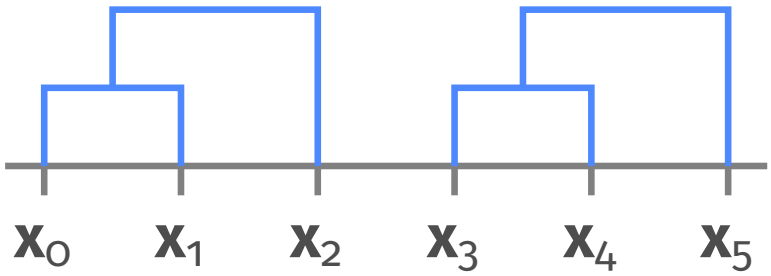


Dendrogram

- Jeder Datenpunkt ist anfangs ein Cluster
- In jedem Schritt werden die nächstgelegenen Cluster zusammengefasst
- Erzeugt Hierarchie von Aufteilungen der Daten

Hierarchisches Clustering (agglomerativ)

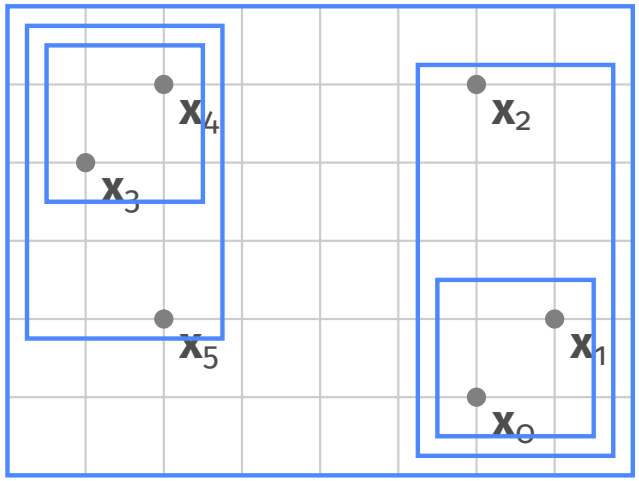


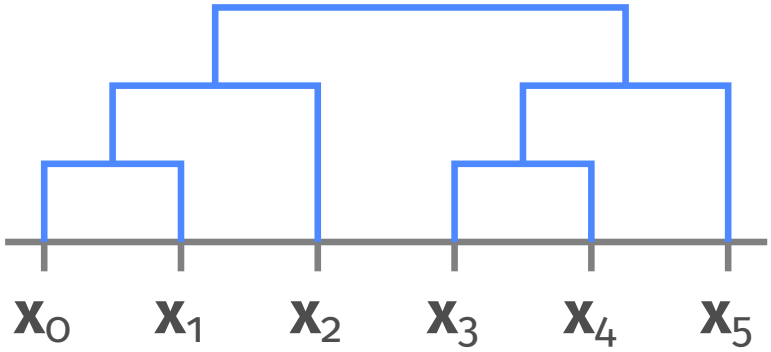


Dendrogram

- Jeder Datenpunkt ist anfangs ein Cluster
- In jedem Schritt werden die nächstgelegenen Cluster zusammengefasst
- Erzeugt Hierarchie von Aufteilungen der Daten

Hierarchisches Clustering (agglomerativ)





Dendrogram

- Jeder Datenpunkt ist anfangs ein Cluster
- In jedem Schritt werden die nächstgelegenen Cluster zusammengefasst
- Erzeugt Hierarchie von Aufteilungen der Daten

k-Means Algorithmus

- Distanz-basiertes, iteratives Clustering-Verfahren
- Erzeugt k disjunkte Teilmengen von $\mathbf{X}$
- k ist Benutzer-Parameter
- Distanzfunktion wird auch von Benutzer gewählt

Algorithmus: k-Means

1. Wähle k zufällige Clusterpunkte $c_1, \ldots, c_k$ aus $\mathbf{X}$
2. Ordne jedes $\mathbf{x} \in \mathbf{X}$ dem nächstgelegenen c_i zu, d.h.

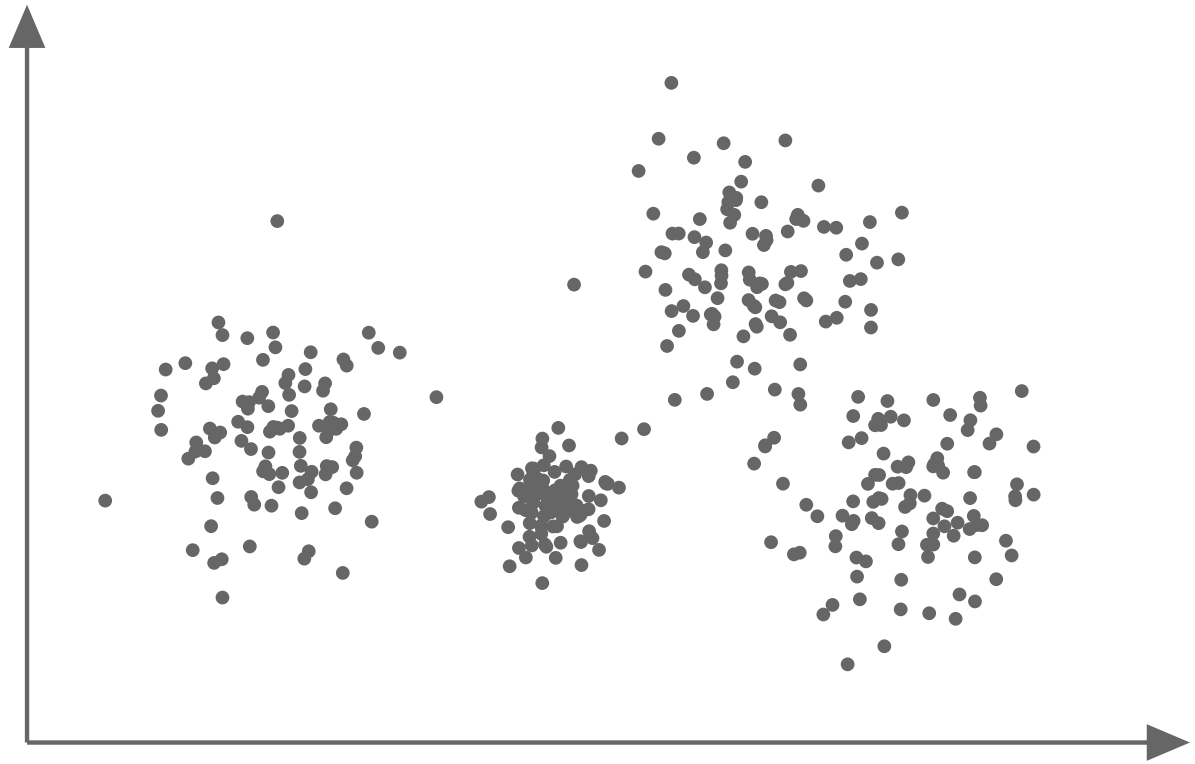
$$\mathbf{C}_i = \{ \mathbf{x} \in \mathbf{X} \mid \mathbf{x} \text{ am nächsten an } \mathbf{c}_i \}$$

3. Berechne neue Clusterpunkte

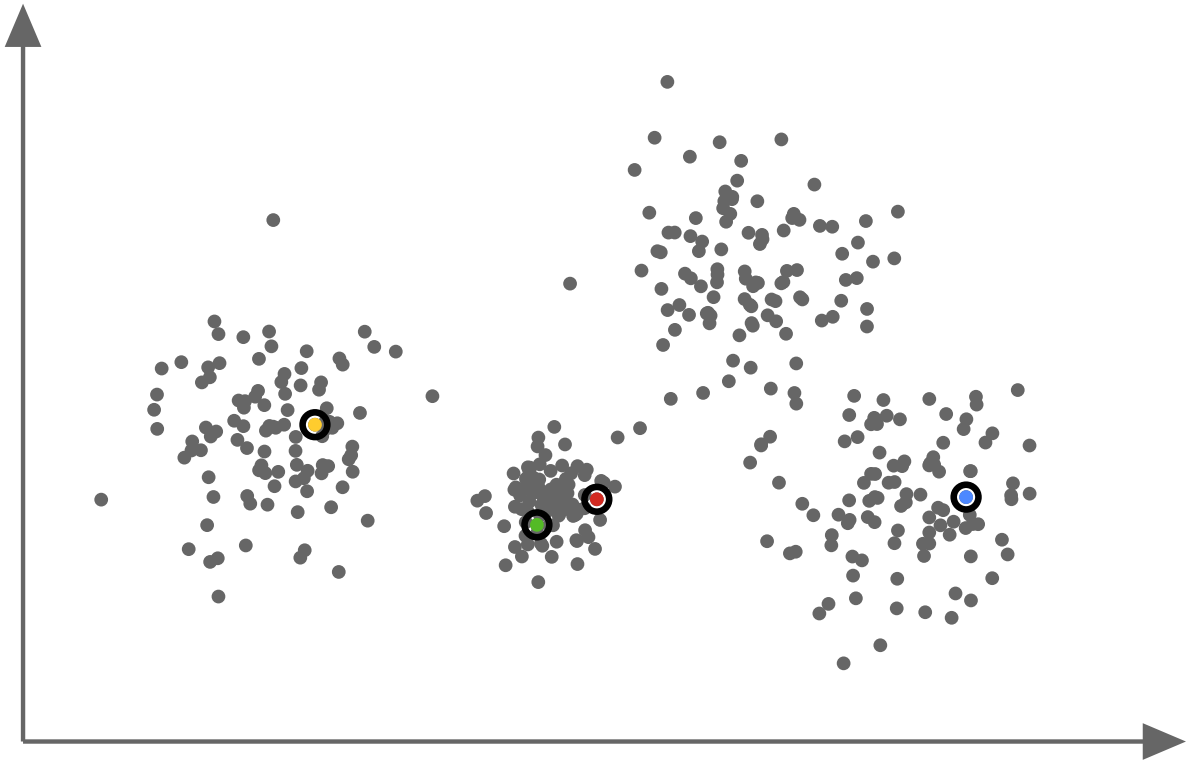
$$\bar{\mathbf{c}}_i = \frac{1}{|\mathbf{C}_i|} \sum_{\mathbf{x_j} \in \mathbf{C}_i} \mathbf{x_j}$$

Wenn $\bar{\mathbf{c}}_1, \ldots, \bar{\mathbf{c}}_k \simeq \mathbf{c}_1, \ldots, \mathbf{c}_k$, dann STOP, sonst springe zu 2. mit den neuen Punkten $\bar{\mathbf{c}}_1, \ldots, \bar{\mathbf{c}}_k$

Beispiel: k-Means

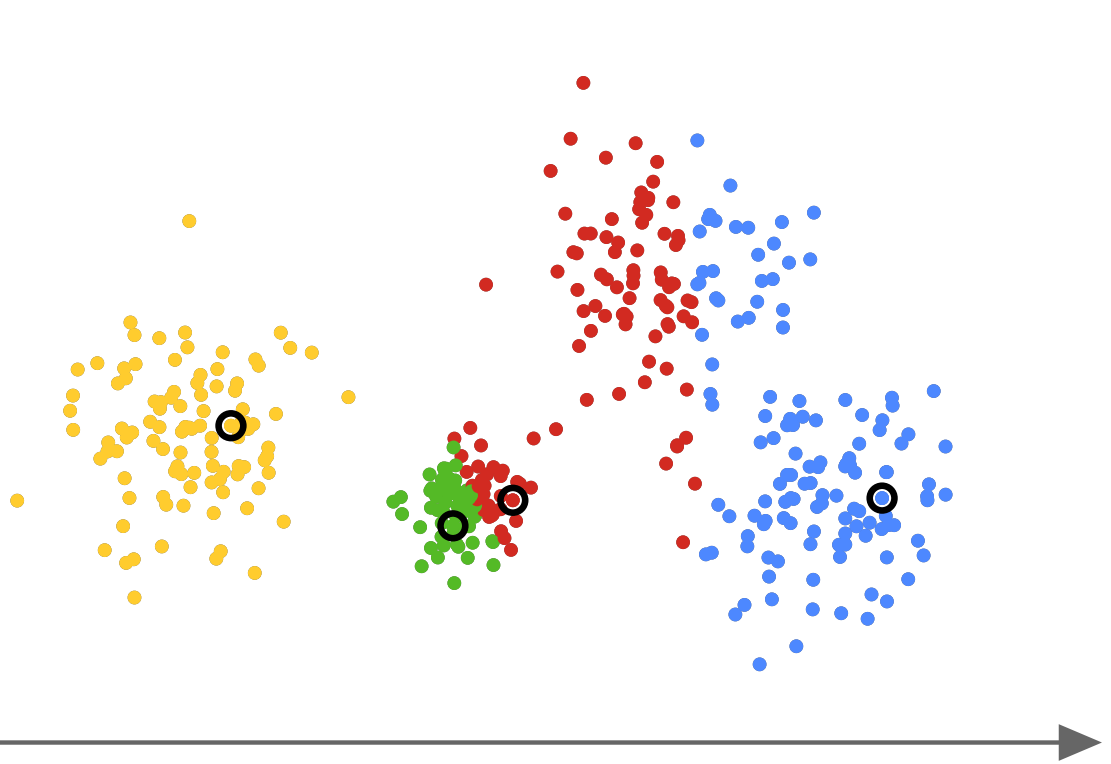


Beispiel: k-Means



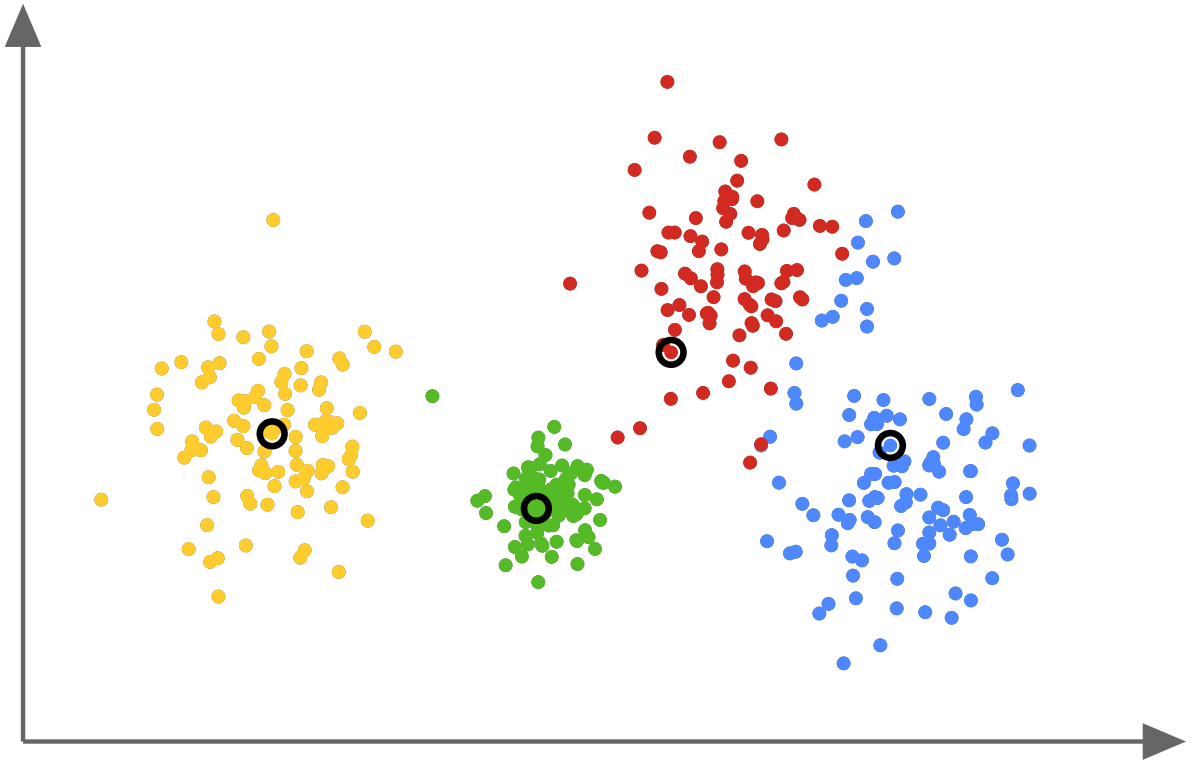
1. Wähle k zufällige Cluster-Mittelpunkte $\mathbf{c}_1, \ldots, \mathbf{c}_k$

Beispiel: k-Means



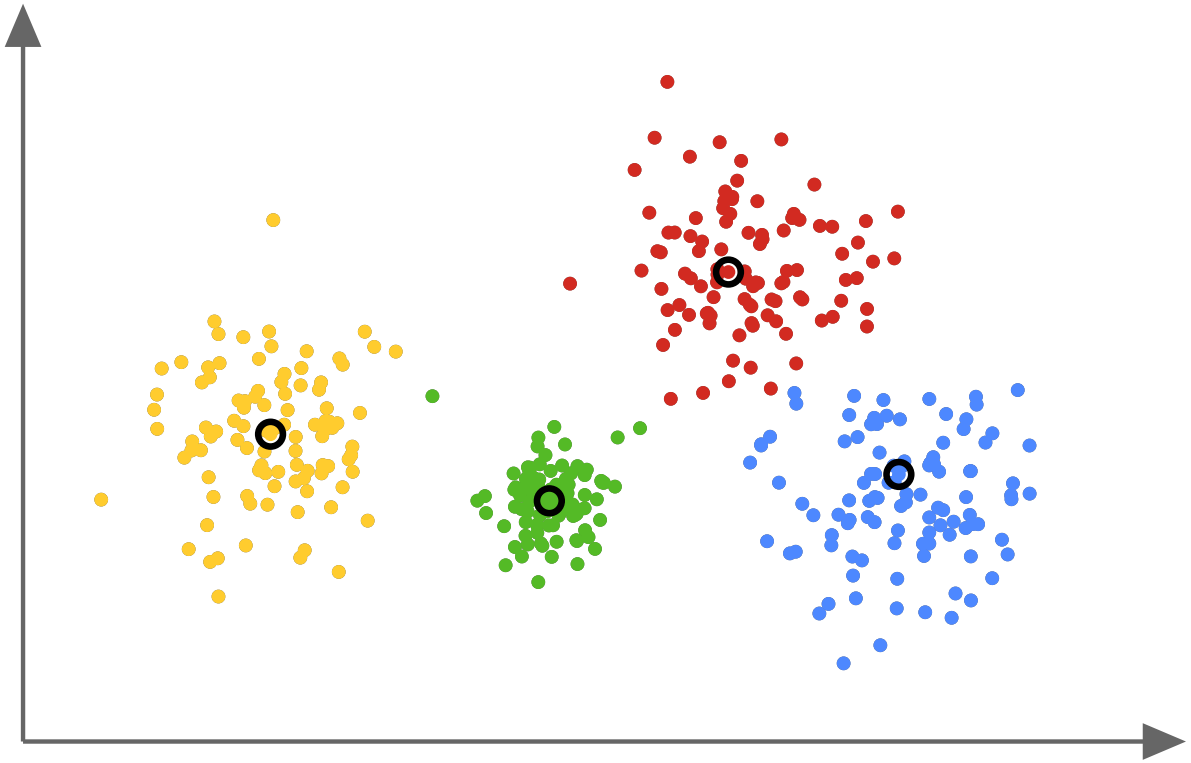
2. Ordne jedem Punkt seinen nächsten Cluster-Mittelpunkt zu

Beispiel: k-Means



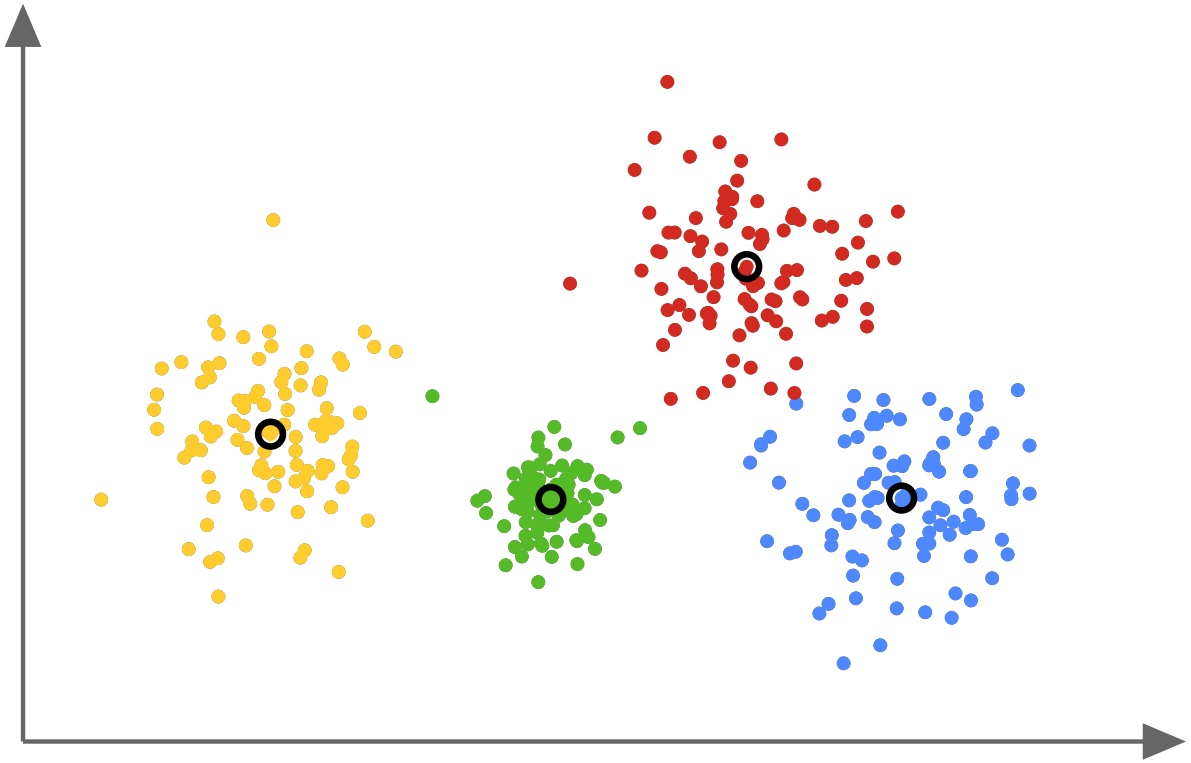
3. Für die Cluster neue Mittelpunkte berechnen, Punkte zuordnen

Beispiel: k-Means



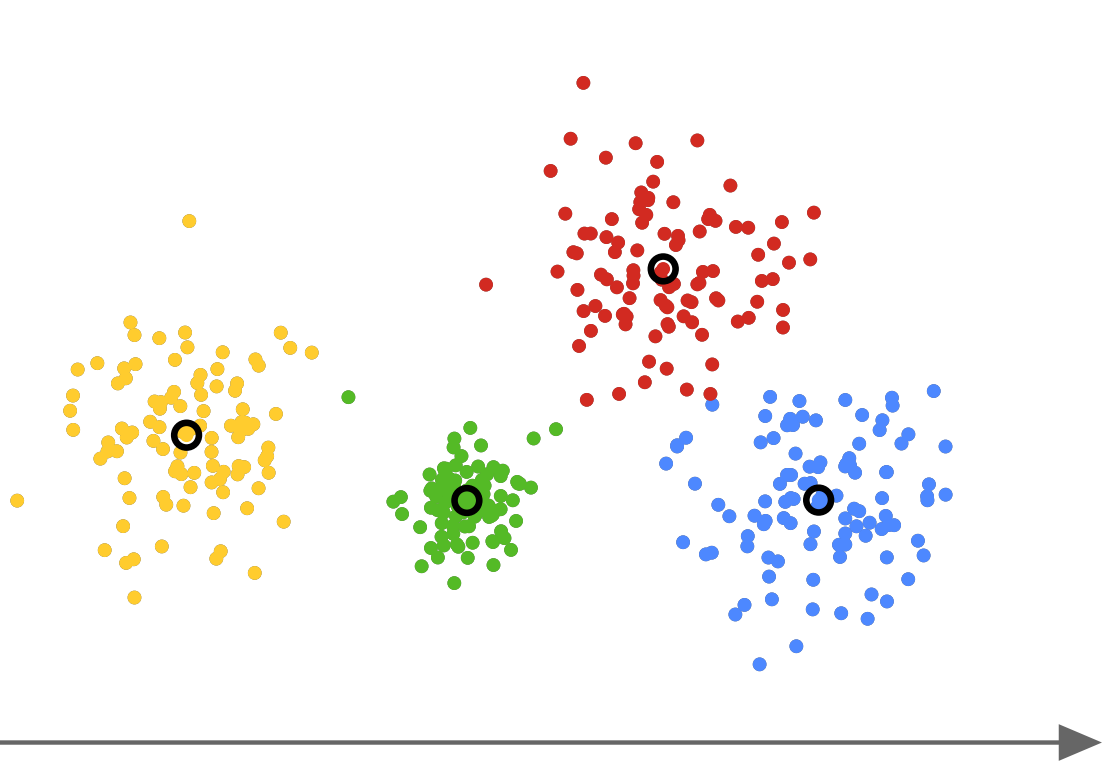
Schritte 2. und 3. wiederholen bis Cluster-Mittelpunkte stabil

Beispiel: k-Means



Schritte 2. und 3. wiederholen bis Cluster-Mittelpunkte stabil

Beispiel: k-Means



Schritte 2. und 3. wiederholen bis Cluster-Mittelpunkte stabil

Clustering - e-Commerce

Clustering von Produkten

- Produktbeschreibungen als Dokument
- Ähnlichkeit von Produkten über Beschreibungen

Ziel: Finde Gruppen ähnlicher Produkte!

Clustering von Produkten

- Produktbeschreibungen als Dokument
- Ähnlichkeit von Produkten über Beschreibungen

Ziel: Finde Gruppen ähnlicher Produkte!

Anwendungen:

- Empfehlen ähnlicher Produkte
- Replacement-Produkte für ausverkaufte Artikel finden!

Kundenprofile – e-Commerce

- Kunden-Daten (Geschlecht, Alter, Wohnort,...)
- Jeder Kunde hat mehrere Produkte gekauft

Kundenprofile – e-Commerce

- Kunden-Daten (Geschlecht, Alter, Wohnort,...)
- Jeder Kunde hat mehrere Produkte gekauft
- Kunde = Dokument mit Worten aus allen gekauften Produkten

Ziel: Finde Gruppen ähnlicher Kunden!

Kundenprofile – e-Commerce

- Kunden-Daten (Geschlecht, Alter, Wohnort,...)
- Jeder Kunde hat mehrere Produkte gekauft
- Kunde = Dokument mit Worten aus allen gekauften Produkten

Ziel: Finde Gruppen ähnlicher Kunden!

Anwendungen:

- Gruppen für gemeinsame Marketing-Maßnahmen
- Bilde Gruppen für Suchprofile (Ranking)

Vorschau: Vorlesung 5

- Warenkorbanalyse