

## K-Means-Clustering

Im Rahmen der Proseminar-Vortragsreihe  
"Grundlagen des Data-Minings für strukturierte Daten"  
Dr. Nils M. Kriege

Antonie Vietor

5. Februar 2018

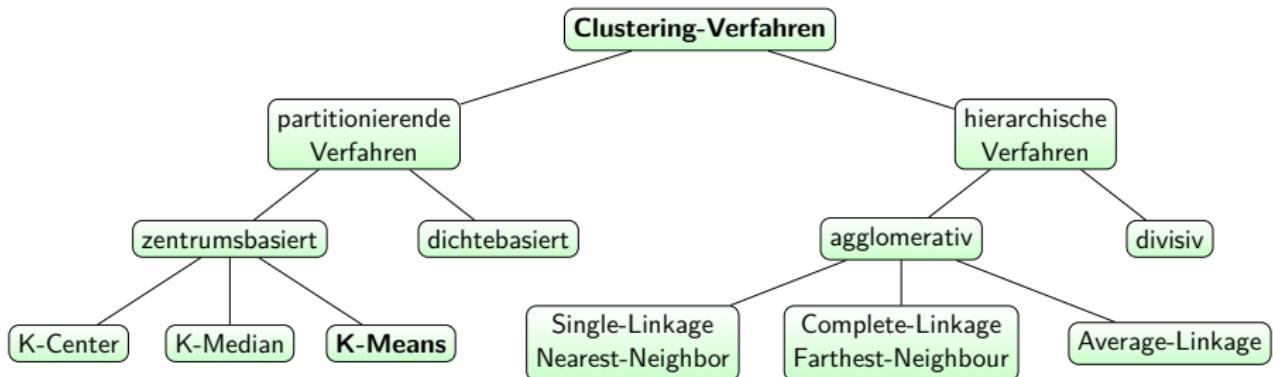
## Überwachtes Lernen

- die **Klassen** der Daten sind bereits **bekannt**
- Ziel: Klassifikationsregeln aufstellen, um neue Daten in die Klassen einordnen zu können

## Unüberwachtes Lernen

- die **Klassen** der Daten sind **nicht bekannt**
- Ziel: Klassen mit möglichst ähnlichen Objekten sollen identifiziert werden

# Clusteringverfahren - Übersicht



modifiziert nach (Backhaus et al., 2008)

# K-Means

## Ziel

Für alle Cluster soll die Summe der quadratischen Abstände der Datenpunkte zu ihrem Clusterzentrum minimiert werden.

## Voraussetzungen

- Anzahl der Klassen muss im Voraus bekannt sein
- Daten stammen aus  $d$ -dimensionalen Raum
- Datensatz besteht aus sphärischen, gut abgrenzbaren Datenpunktanhäufungen

# K-Means

Definition (Summe der quadratischen Abstände **eines** Clusters)

$$J(c_j) = \sum_{a_i \in c_j} \|a_i - c_j\|^2 = \sum_{a_i \in c_j} d^2(a_i, c_j)$$

$n$

$C = \{c_1, \dots, c_k\}$

$A = \{a_1, \dots, a_n\}$

Anzahl der Datenpunkte

Menge der Vektoren der Clusterzentren

Menge der Datenpunkte

# K-Means

Definition (Summe der quadratischen Abstände **eines** Clusters)

$$J(c_j) = \sum_{a_i \in c_j} \|a_i - c_j\|^2 = \sum_{a_i \in c_j} d^2(a_i, c_j)$$

Definition (Summe der quadratischen Abstände **aller** Cluster)

$$J(C) = \sum_{j=1}^k J(c_j) = \sum_{j=1}^k \sum_{a_i \in c_j} d^2(a_i, c_j)$$

$k$

Clusteranzahl

$n$

Anzahl der Datenpunkte

$C = \{c_1, \dots, c_k\}$

Menge der Vektoren der Clusterzentren

$A = \{a_1, \dots, a_n\}$

Menge der Datenpunkte

# K-Means-Basisalgorithmus nach Lloyd

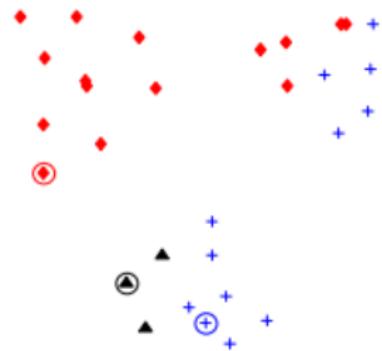
1: Wähle  $k$  Clusterzentren.



Bildquelle: Jain (2010)

# K-Means-Basisalgorithmus nach Lloyd

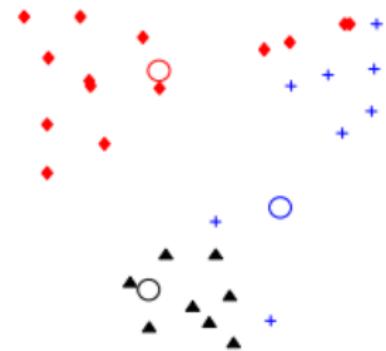
- 1: Wähle  $k$  Clusterzentren.
- 2: Ordne jeden Datenpunkt dem nächstliegenden Clusterzentrum zu.



Bildquelle: Jain (2010)

# K-Means-Basisalgorithmus nach Lloyd

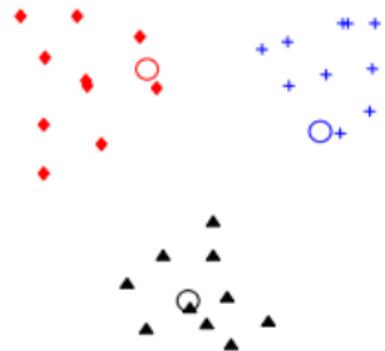
- 1: Wähle  $k$  Clusterzentren.
- 2: Ordne jeden Datenpunkt dem nächstliegenden Clusterzentrum zu.
- 3: Bestimme das Clusterzentrum jedes Clusters erneut und ersetze die bisherigen Clusterzentren.



Bildquelle: Jain (2010)

# K-Means-Basisalgorithmus nach Lloyd

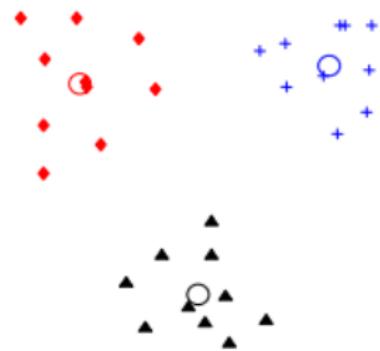
- 1: Wähle  $k$  Clusterzentren.
- 2: Ordne jeden Datenpunkt dem nächstliegenden Clusterzentrum zu.
- 3: Bestimme das Clusterzentrum jedes Clusters erneut und ersetze die bisherigen Clusterzentren.
- 4: Wiederhole Schritt 2 und 3 bis sich die Positionen der Clusterzentren nicht mehr ändern.



Bildquelle: Jain (2010)

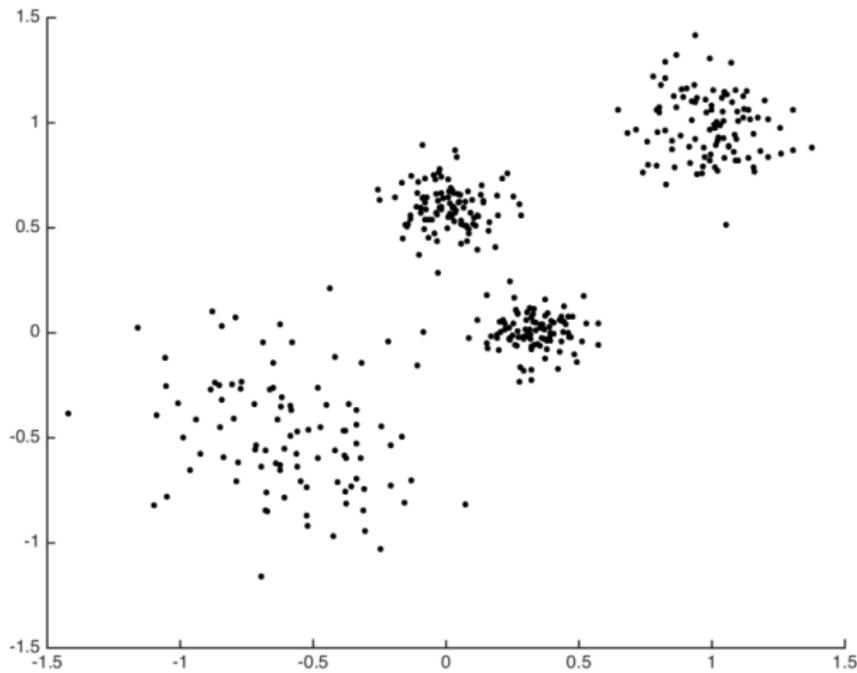
# K-Means-Basisalgorithmus nach Lloyd

- 1: Wähle  $k$  Clusterzentren.
- 2: Ordne jeden Datenpunkt dem nächstliegenden Clusterzentrum zu.
- 3: Bestimme das Clusterzentrum jedes Clusters erneut und ersetze die bisherigen Clusterzentren.
- 4: Wiederhole Schritt 2 und 3 bis sich die Positionen der Clusterzentren nicht mehr ändern.



Bildquelle: Jain (2010)

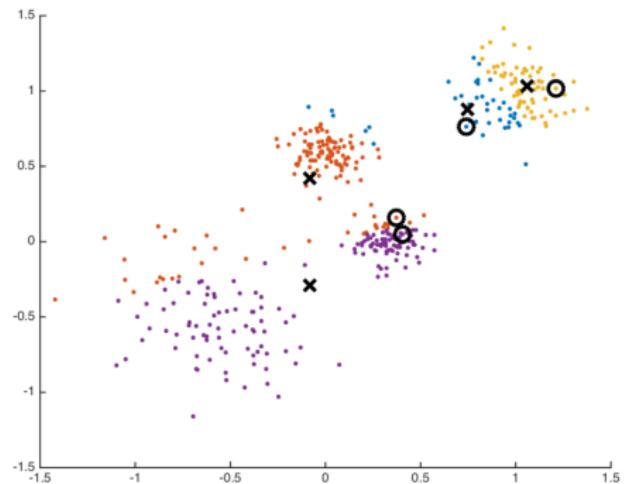
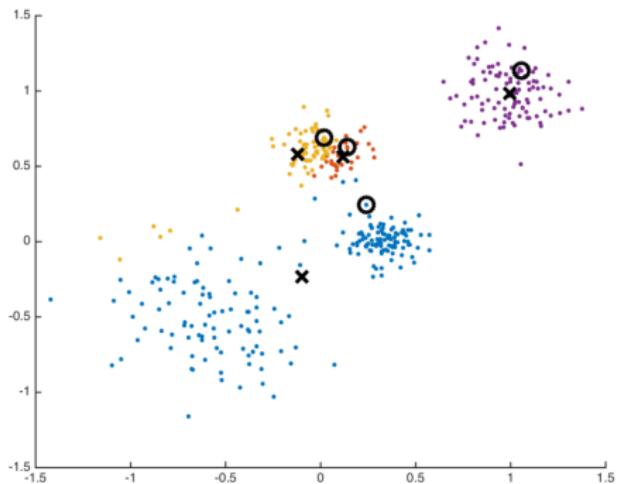
# Lokales vs. globales Optimum



Bildquelle: Stotz (2016)

# Lokales vs. globales Optimum

## 1. Iteration



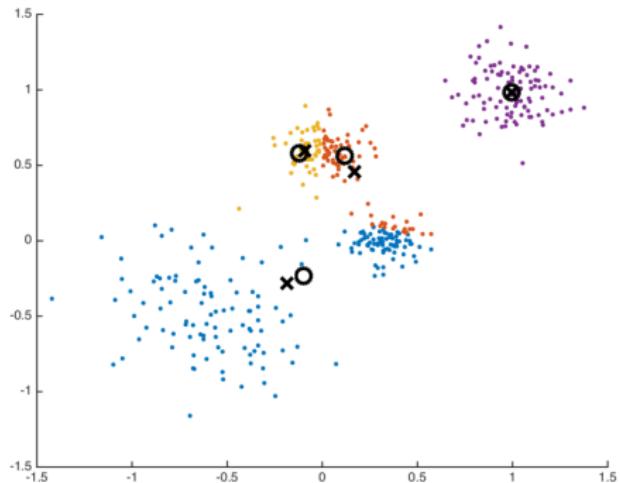
**o** = Clusterzentren zu Beginn der Iteration

**x** = Clusterzentren am Ende der Iteration

Bildquelle: Stotz (2016)

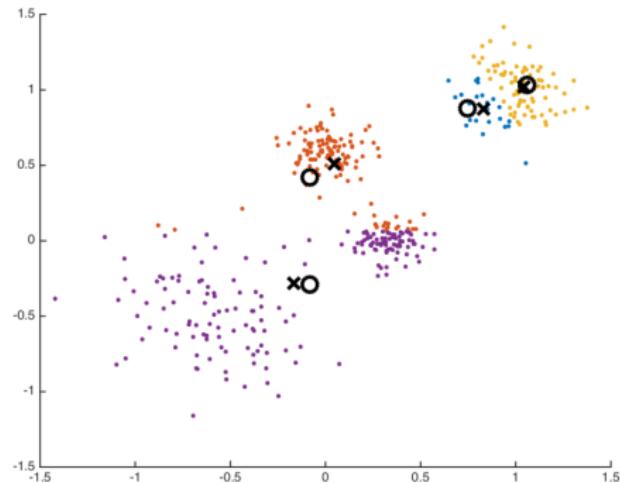
# Lokales vs. globales Optimum

## 2. Iteration



○ = Clusterzentren zu Beginn der Iteration

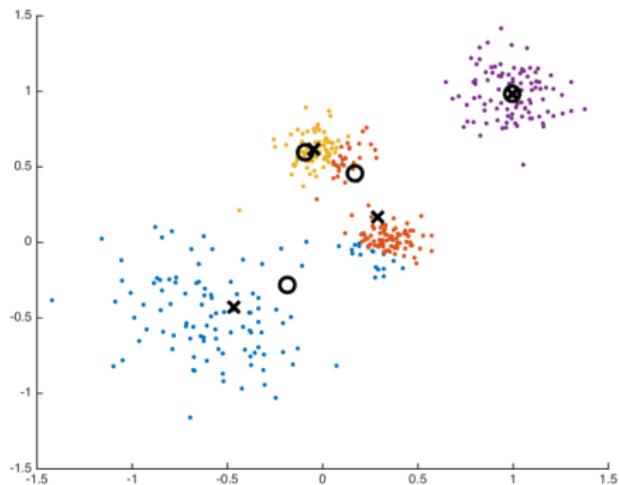
✗ = Clusterzentren am Ende der Iteration



Bildquelle: Stotz (2016)

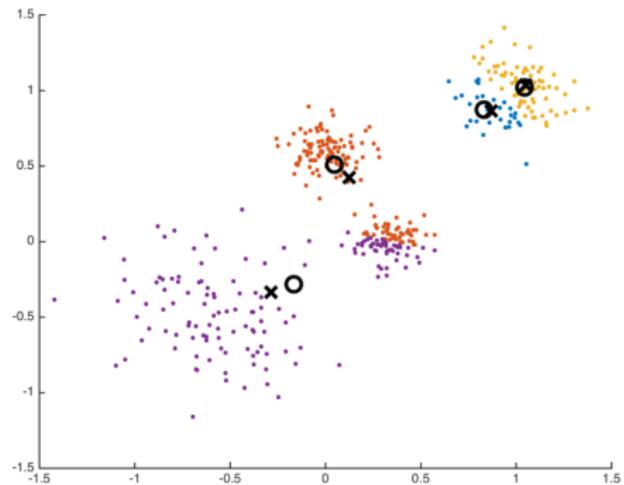
# Lokales vs. globales Optimum

## 3. Iteration



○ = Clusterzentren zu Beginn der Iteration

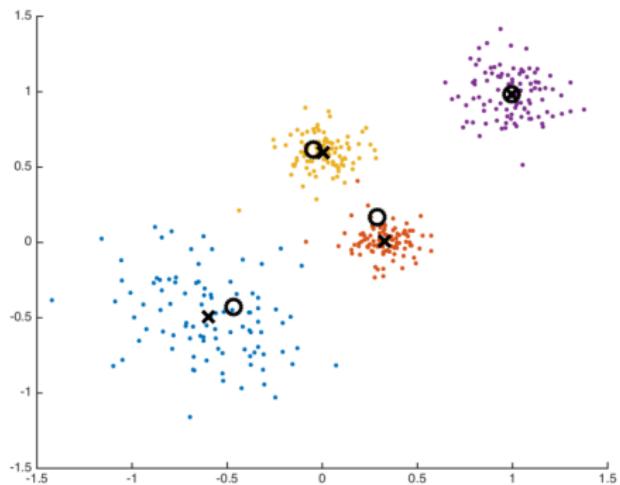
✗ = Clusterzentren am Ende der Iteration



Bildquelle: Stotz (2016)

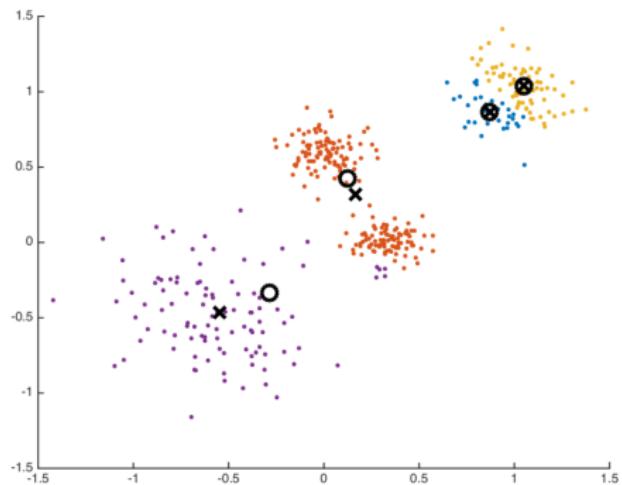
# Lokales vs. globales Optimum

## 4. Iteration



**o** = Clusterzentren zu Beginn der Iteration

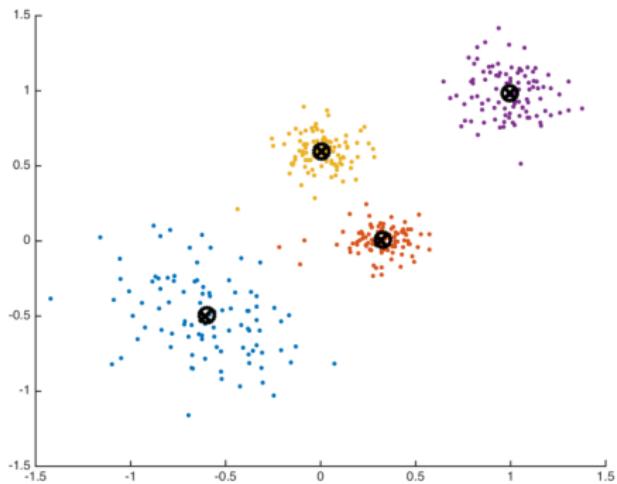
**x** = Clusterzentren am Ende der Iteration



Bildquelle: Stotz (2016)

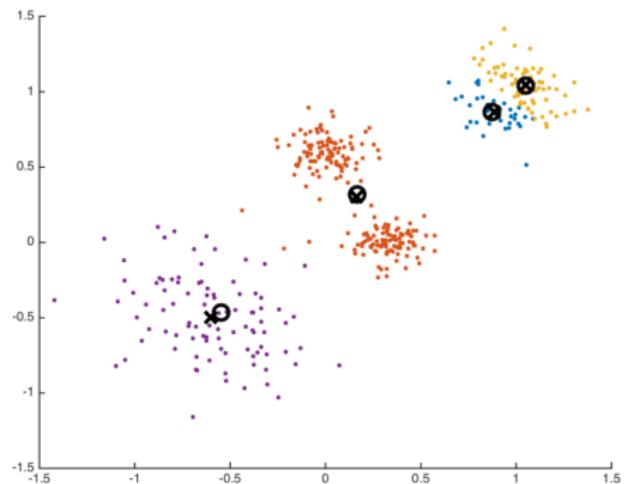
# Lokales vs. globales Optimum

## 5. Iteration



o = Clusterzentren zu Beginn der Iteration

x = Clusterzentren am Ende der Iteration



Bildquelle: Stotz (2016)

# Die optimale Clusteranzahl $k$ bestimmen

## Clusterradius

Maximale Distanz zwischen den Punkten eines Clusters und dem Clusterzentrum.

## Clusterdurchmesser

Maximale Distanz zwischen zwei Punkten eines Clusters.

- Solange  $k$  der tatsächlichen Clusteranzahl entspricht, verändern sich Clusterradius bzw. -durchmesser nur geringfügig beim Hinzufügen von weiteren Punkten.
- Werden zwei Cluster zusammengefasst, die nicht zueinander passen, dann kommt es zu einem deutlichen Anstieg des Clusterradius bzw. -durchmessers.

# Die optimale Clusteranzahl $k$ bestimmen

## Clusterradius

Maximale Distanz zwischen den Punkten eines Clusters und dem Clusterzentrum.

## Clusterdurchmesser

Maximale Distanz zwischen zwei Punkten eines Clusters.

- Solange  $k$  der tatsächlichen Clusteranzahl entspricht, verändern sich Clusterradius bzw. -durchmesser nur geringfügig beim Hinzufügen von weiteren Punkten.
- Werden zwei Cluster zusammengefasst, die nicht zueinander passen, dann kommt es zu einem deutlichen Anstieg des Clusterradius bzw. -durchmessers.

# Die optimale Clusteranzahl $k$ bestimmen

## Clusterradius

Maximale Distanz zwischen den Punkten eines Clusters und dem Clusterzentrum.

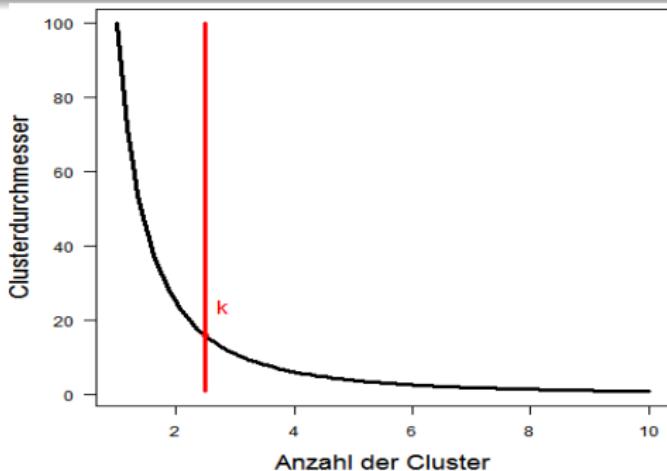
## Clusterdurchmesser

Maximale Distanz zwischen zwei Punkten eines Clusters.

- Solange  $k$  der tatsächlichen Clusteranzahl entspricht, verändern sich Clusterradius bzw. -durchmesser nur geringfügig beim Hinzufügen von weiteren Punkten.
- Werden zwei Cluster zusammengefasst, die nicht zueinander passen, dann kommt es zu einem deutlichen Anstieg des Clusterradius bzw. -durchmessers.

# Die optimale Clusteranzahl $k$ bestimmen

- wiederholte Messungen unter Beobachtung von Clusterradius oder -durchmesser
- effektiv ist die Wahl von Zweierpotenzen für  $k$  bis sich Clusterradius oder -durchschnitt nur noch geringfügig verändern
- weitere Eingrenzung des Intervalls kann dann durch binäre Aufteilungen (binäre Suche) erfolgen



# BFR-Algorithmus

(nach Bradley, Fayyad und Reina)

## Eigenschaften

- für multivariate Daten geeignet
- für Big Data geeignet

## Voraussetzungen

- die Daten müssen aus dem **mehrdimensionalen** Raum stammen
- in jeder Dimension **Normalverteilung** der Datenpunkte um die Clusterzentren herum

# Initialisierung

- 1 initiale Auswahl einer optimalen Anzahl von  $k$  Punkten

# Initialisierung

- ① initiale Auswahl einer optimalen Anzahl von  $k$  Punkten
- ② zufällige Aufteilung der Daten in mehrere Stichproben

# Initialisierung

- ① initiale Auswahl einer optimalen Anzahl von  $k$  Punkten
- ② zufällige Aufteilung der Daten in mehrere Stichproben
- ③ anlegen von drei Daten-Sets:

# Initialisierung

- ① initiale Auswahl einer optimalen Anzahl von  $k$  Punkten
- ② zufällige Aufteilung der Daten in mehrere Stichproben
- ③ anlegen von drei Daten-Sets:
  - Discard Set (**DS**)
  - Compressed Set (**CS**)
  - Retained Set (**RS**)

# Daten-Sets

- **DS (Discard Set):**

⇒ repräsentiert die eigentlichen **Cluster**

Für Punkte, die nahe genug an einem der  $k$  Clusterzentren liegen.

# Daten-Sets

- **DS (Discard Set):**

⇒ repräsentiert die eigentlichen **Cluster**

Für Punkte, die nahe genug an einem der  $k$  Clusterzentren liegen.

- **CS (Compressed Set):**

⇒ repräsentiert zusätzliche **Minicluster**

Für Gruppen von nahe bei einander liegenden Punkten, die nicht nahe genug an einem der Clusterzentren liegen, um diesem zugeordnet zu werden.

# Daten-Sets

- **DS (Discard Set):**

⇒ repräsentiert die eigentlichen **Cluster**

Für Punkte, die nahe genug an einem der  $k$  Clusterzentren liegen.

- **CS (Compressed Set):**

⇒ repräsentiert zusätzliche **Minicluster**

Für Gruppen von nahe bei einander liegenden Punkten, die nicht nahe genug an einem der Clusterzentren liegen, um diesem zugeordnet zu werden.

- **RS (Retained Set):**

⇒ repräsentiert **isolierte Datenpunkte**

Für Datenpunkte, die weder nahe genug an den Clusterzentren des CS noch an denen des DS liegen.

# Datenanalyse

Für jedes DS und CS werden folgende Daten gespeichert:

- $N$ : Anzahl der Datenpunkte

# Datenanalyse

Für jedes DS und CS werden folgende Daten gespeichert:

- **$N$** : Anzahl der Datenpunkte
- **$SUM$** : Vektor, dessen Elemente jeweils die Summe der Koordinaten einer Dimension von allen Punkten des Clusters repräsentieren.

# Datenanalyse

Für jedes DS und CS werden folgende Daten gespeichert:

- **$N$** : Anzahl der Datenpunkte
- **$SUM$** : Vektor, dessen Elemente jeweils die Summe der Koordinaten einer Dimension von allen Punkten des Clusters repräsentieren.
- **$SUMSQ$** : Vektor, dessen Elemente jeweils die Summe der Koordinatenquadrate einer Dimension von allen Punkten des Clusters repräsentieren.

# Datenanalyse

Für jedes DS und CS werden folgende Daten gespeichert:

- **$N$** : Anzahl der Datenpunkte
- **$SUM$** : Vektor, dessen Elemente jeweils die Summe der Koordinaten einer Dimension von allen Punkten des Clusters repräsentieren.
- **$SUMSQ$** : Vektor, dessen Elemente jeweils die Summe der Koordinatenquadrate einer Dimension von allen Punkten des Clusters repräsentieren.

**Beispiel:**

$$a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

# Datenanalyse

Für jedes DS und CS werden folgende Daten gespeichert:

- **$N$** : Anzahl der Datenpunkte
- **$SUM$** : Vektor, dessen Elemente jeweils die Summe der Koordinaten einer Dimension von allen Punkten des Clusters repräsentieren.
- **$SUMSQ$** : Vektor, dessen Elemente jeweils die Summe der Koordinatenquadrate einer Dimension von allen Punkten des Clusters repräsentieren.

**Beispiel:**  $\Rightarrow N = 2$

$$a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

# Datenanalyse

Für jedes DS und CS werden folgende Daten gespeichert:

- **$N$** : Anzahl der Datenpunkte
- **$SUM$** : Vektor, dessen Elemente jeweils die Summe der Koordinaten einer Dimension von allen Punkten des Clusters repräsentieren.
- **$SUMSQ$** : Vektor, dessen Elemente jeweils die Summe der Koordinatenquadrate einer Dimension von allen Punkten des Clusters repräsentieren.

Beispiel:

$$\Rightarrow N = 2$$

$$a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

$$\Rightarrow SUM = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \end{pmatrix}$$

# Datenanalyse

Für jedes DS und CS werden folgende Daten gespeichert:

- **$N$** : Anzahl der Datenpunkte
- **$SUM$** : Vektor, dessen Elemente jeweils die Summe der Koordinaten einer Dimension von allen Punkten des Clusters repräsentieren.
- **$SUMSQ$** : Vektor, dessen Elemente jeweils die Summe der Koordinatenquadrate einer Dimension von allen Punkten des Clusters repräsentieren.

**Beispiel:**

$$a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

$$\Rightarrow N = 2$$

$$\Rightarrow SUM = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \end{pmatrix}$$

$$\Rightarrow SUMSQ = \begin{pmatrix} a_1^2 + b_1^2 \\ a_2^2 + b_2^2 \end{pmatrix}$$

# Datenanalyse

Aus  $N$ ,  $SUM$  und  $SUMSQ$  können die Koordinaten des Clusterzentrums und die Standardabweichung berechnet werden:

# Datenanalyse

Aus  $N$ ,  $SUM$  und  $SUMSQ$  können die Koordinaten des Clusterzentrums und die Standardabweichung berechnet werden:

## Clusterzentrum

Die  $i$ -te Koordinate des Clusterzentrums ergibt sich aus den Mittelwerten der Datenpunktkoordinaten:

$$\frac{SUM_i}{N}$$

# Datenanalyse

Aus  $N$ ,  $SUM$  und  $SUMSQ$  können die Koordinaten des Clusterzentrums und die Standardabweichung berechnet werden:

## Clusterzentrum

Die  $i$ -te Koordinate des Clusterzentrums ergibt sich aus den Mittelwerten der Datenpunktkoordinaten:

$$\frac{SUM_i}{N}$$

## Standardabweichung

Die Varianz  $v$  in der  $i$ -ten Dimension kann aus den  $SUM$ - und  $SUMSQ$ -Vektoren ermittelt werden:

$$v_i = \frac{SUMSQ_i}{N} - \left( \frac{SUM_i}{N} \right)^2$$

Daraus ergibt sich die Standardabweichung  $s_i = \sqrt{v_i}$

# Datenanalyse

Warum entsprechen die Koordinaten des Clusterzentrums den Mittelwerten der Datenpunktkoordinaten?

# Datenanalyse

Warum entsprechen die Koordinaten des Clusterzentrums den Mittelwerten der Datenpunktkoordinaten?

Seien die Koordinaten eines Clusterzentrums  $c$  also:

$$c = \frac{1}{n} \sum_{i=1}^n a_i = \bar{a}$$

# Datenanalyse

Warum entsprechen die Koordinaten des Clusterzentrums den Mittelwerten der Datenpunktkoordinaten?

Seien die Koordinaten eines Clusterzentrums  $c$  also:

$$c = \frac{1}{n} \sum_{i=1}^n a_i = \bar{a}$$

Dann gilt für die Summe der quadratischen Abstände aller Punkte  $a_i$  eines Clusters zu einem Punkt  $x$ :

$$\sum_{i=1}^n |a_i - x|^2$$

# Datenanalyse

$$\sum_{i=1}^n |a_i - x|^2$$

# Datenanalyse

$$\sum_{i=1}^n |a_i - x|^2 = \sum_{i=1}^n |a_i - c + c - x|^2$$

# Datenanalyse

$$\begin{aligned}\sum_{i=1}^n |a_i - x|^2 &= \sum_{i=1}^n |a_i - c + c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + 2(c - x) \cdot \sum_{i=1}^n (a_i - c) + \sum_{i=1}^n |c - x|^2\end{aligned}$$

# Datenanalyse

$$\begin{aligned}\sum_{i=1}^n |a_i - x|^2 &= \sum_{i=1}^n |a_i - c + c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + 2(c - x) \cdot \sum_{i=1}^n (a_i - c) + \sum_{i=1}^n |c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + 2(c - x) \cdot \sum_{i=1}^n (a_i - \bar{a}) + n|c - x|^2\end{aligned}$$

# Datenanalyse

$$\begin{aligned}\sum_{i=1}^n |a_i - x|^2 &= \sum_{i=1}^n |a_i - c + c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + 2(c - x) \cdot \sum_{i=1}^n (a_i - c) + \sum_{i=1}^n |c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + 2(c - x) \cdot \sum_{i=1}^n (a_i - \bar{a}) + n|c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + 2(c - x) \cdot 0 + n|c - x|^2\end{aligned}$$

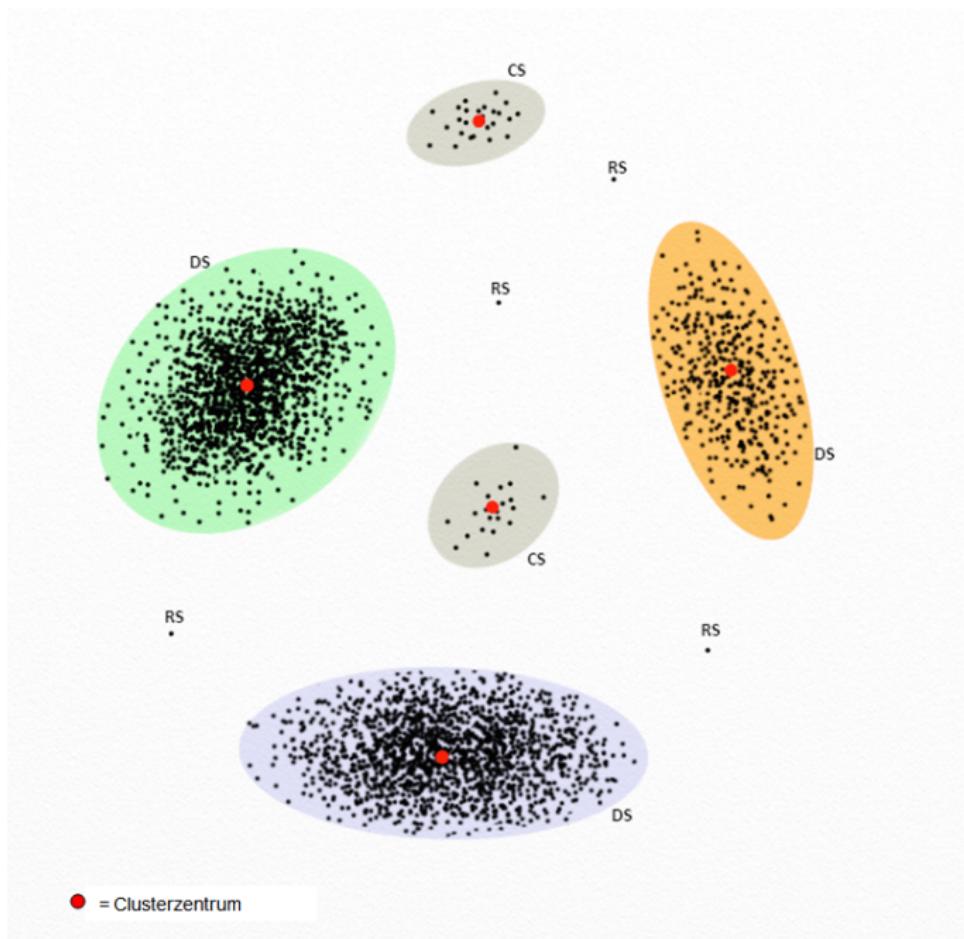
# Datenanalyse

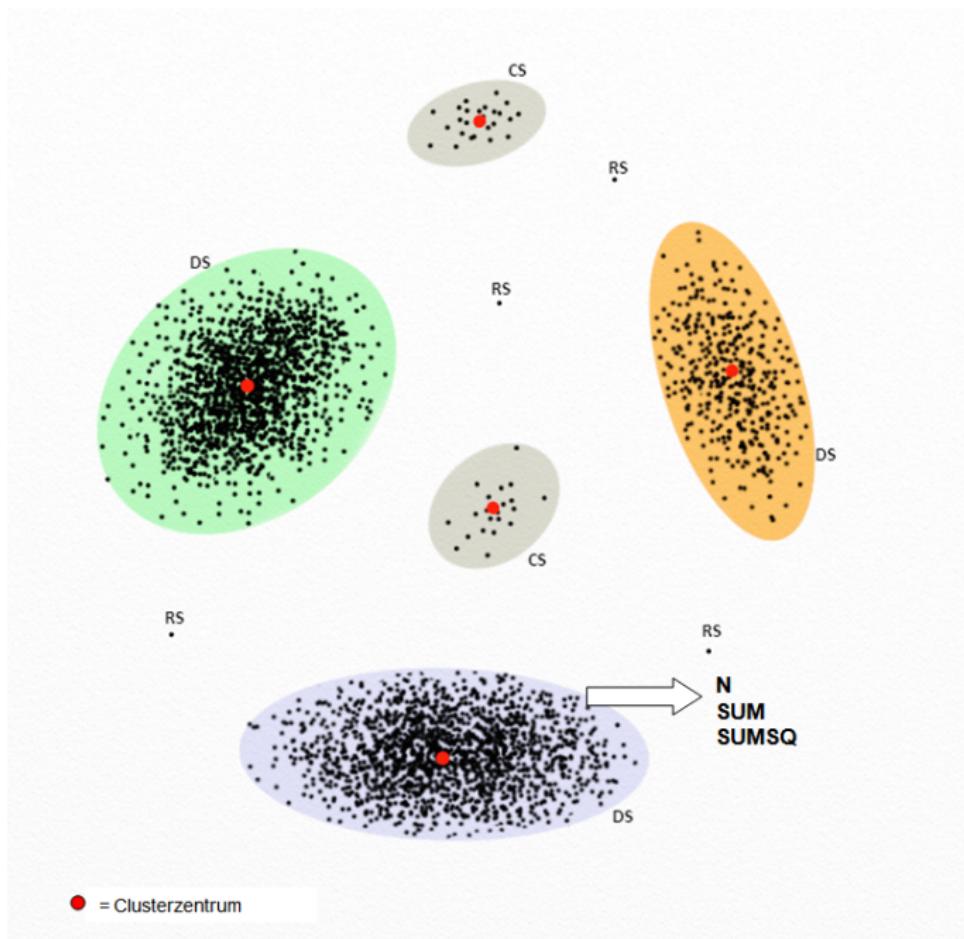
$$\begin{aligned}\sum_{i=1}^n |a_i - x|^2 &= \sum_{i=1}^n |a_i - c + c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + 2(c - x) \cdot \sum_{i=1}^n (a_i - c) + \sum_{i=1}^n |c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + 2(c - x) \cdot \sum_{i=1}^n (a_i - \bar{a}) + n|c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + 2(c - x) \cdot 0 + n|c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + n|c - x|^2\end{aligned}$$

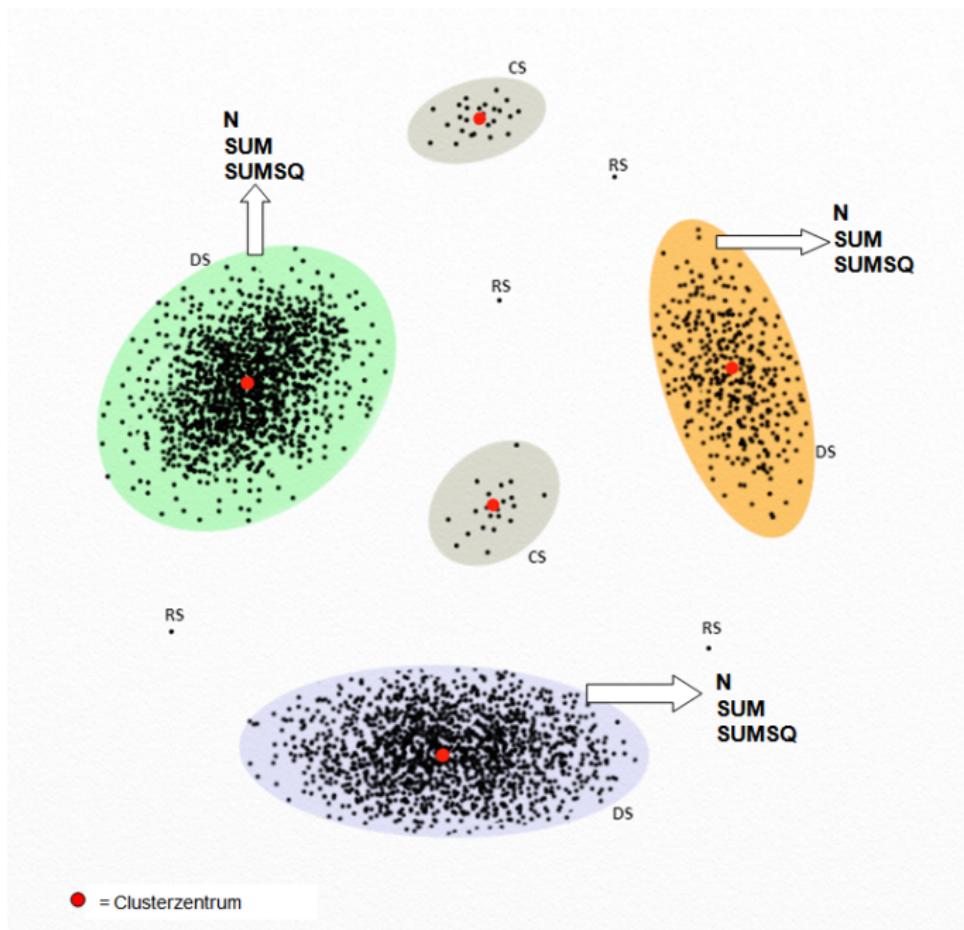
# Datenanalyse

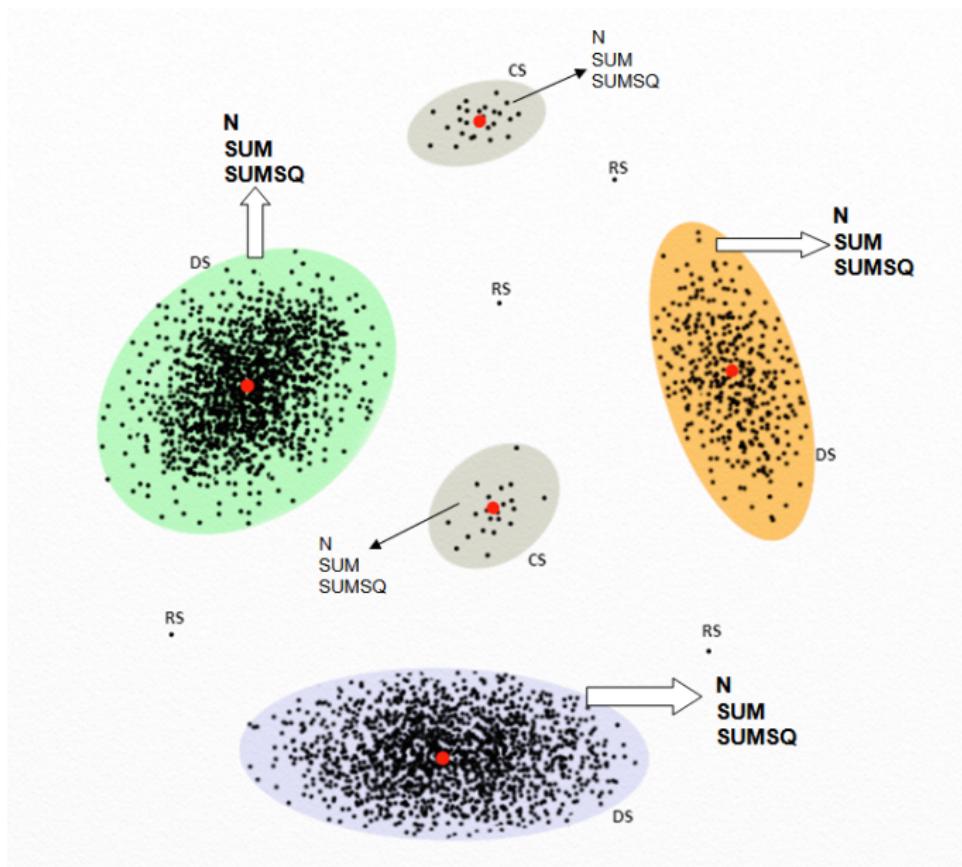
$$\begin{aligned}\sum_{i=1}^n |a_i - x|^2 &= \sum_{i=1}^n |a_i - c + c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + 2(c - x) \cdot \sum_{i=1}^n (a_i - c) + \sum_{i=1}^n |c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + 2(c - x) \cdot \sum_{i=1}^n (a_i - \bar{a}) + n|c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + 2(c - x) \cdot 0 + n|c - x|^2 \\&= \sum_{i=1}^n |a_i - c|^2 + n|c - x|^2\end{aligned}$$

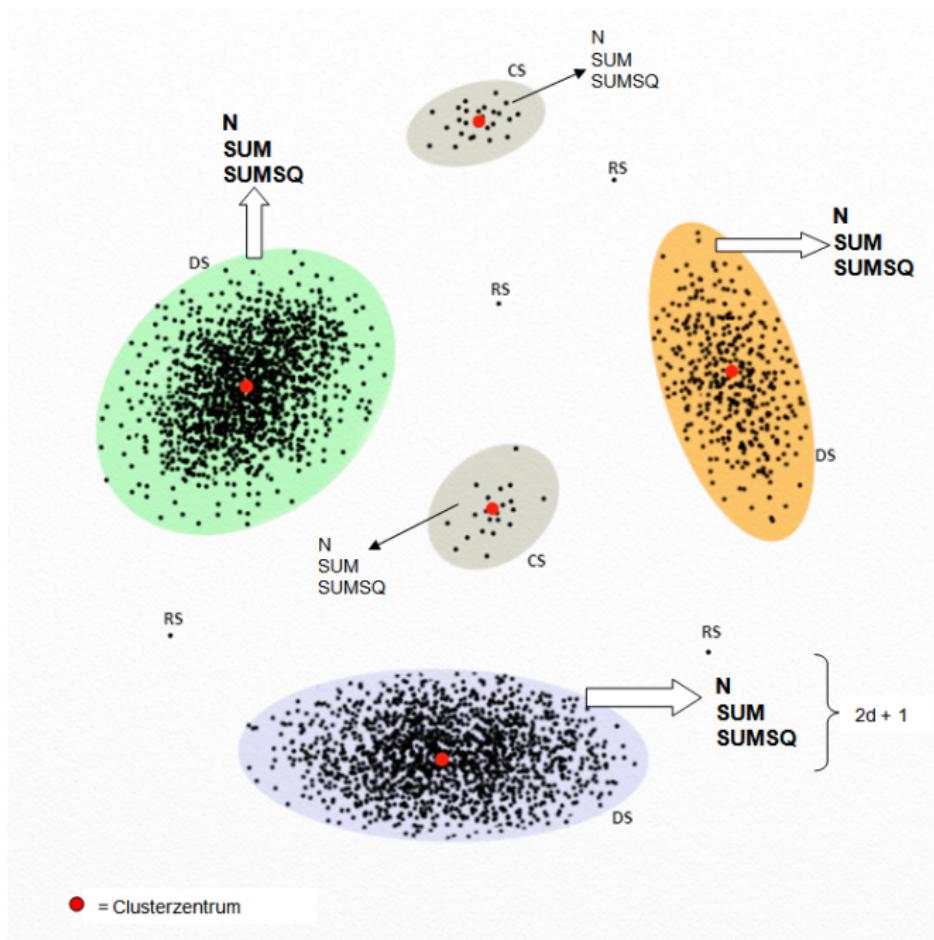
⇒ ist genau dann minimal, wenn  $x = c$  und die Koordinaten von  $c$  den **Mittelwerten** der Datenpunktkoordinaten entsprechen.

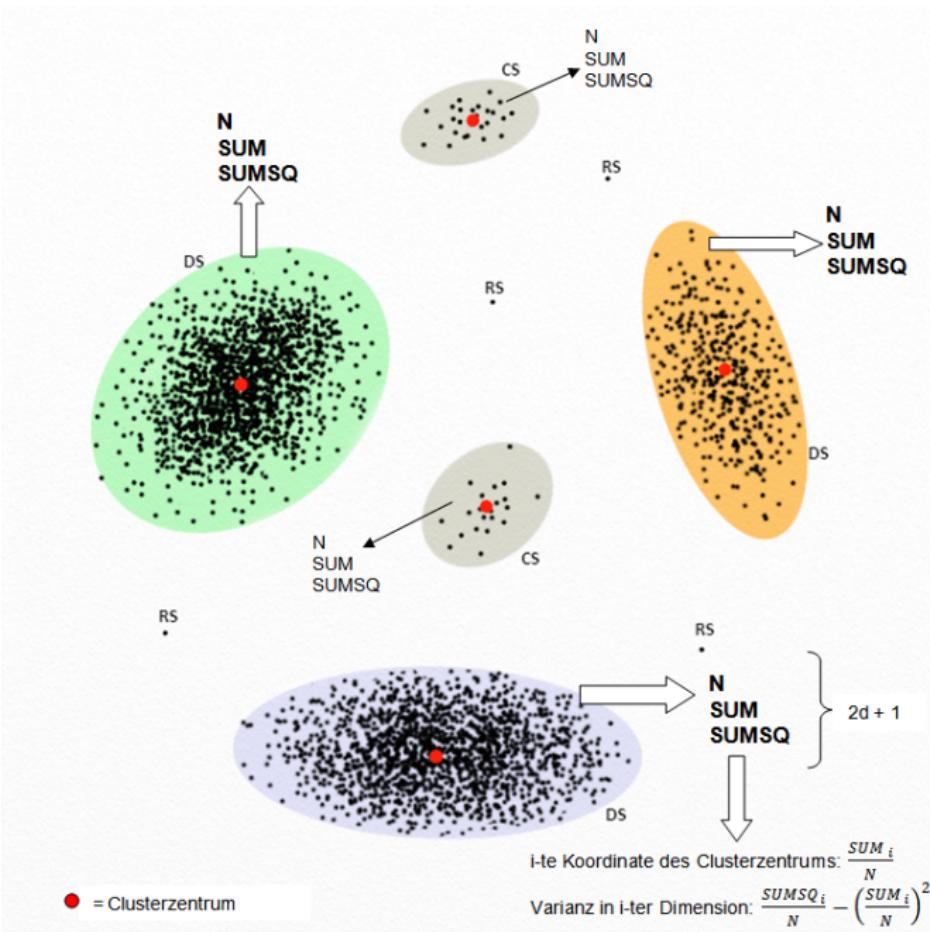


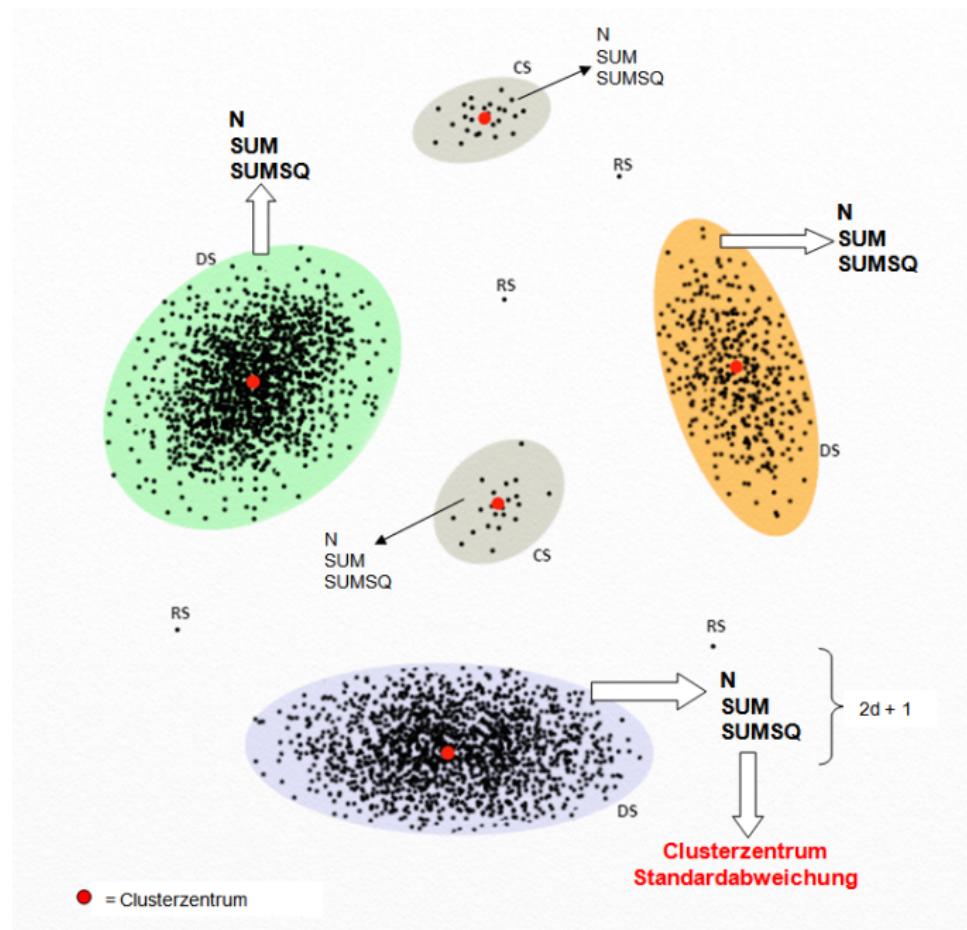


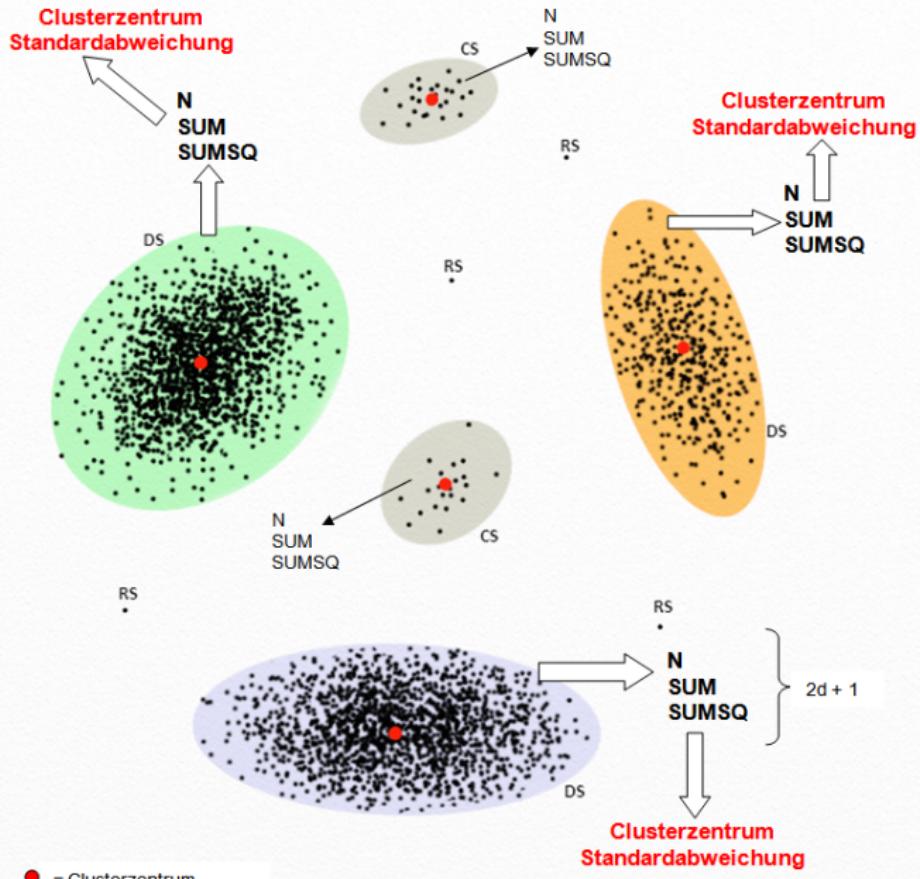


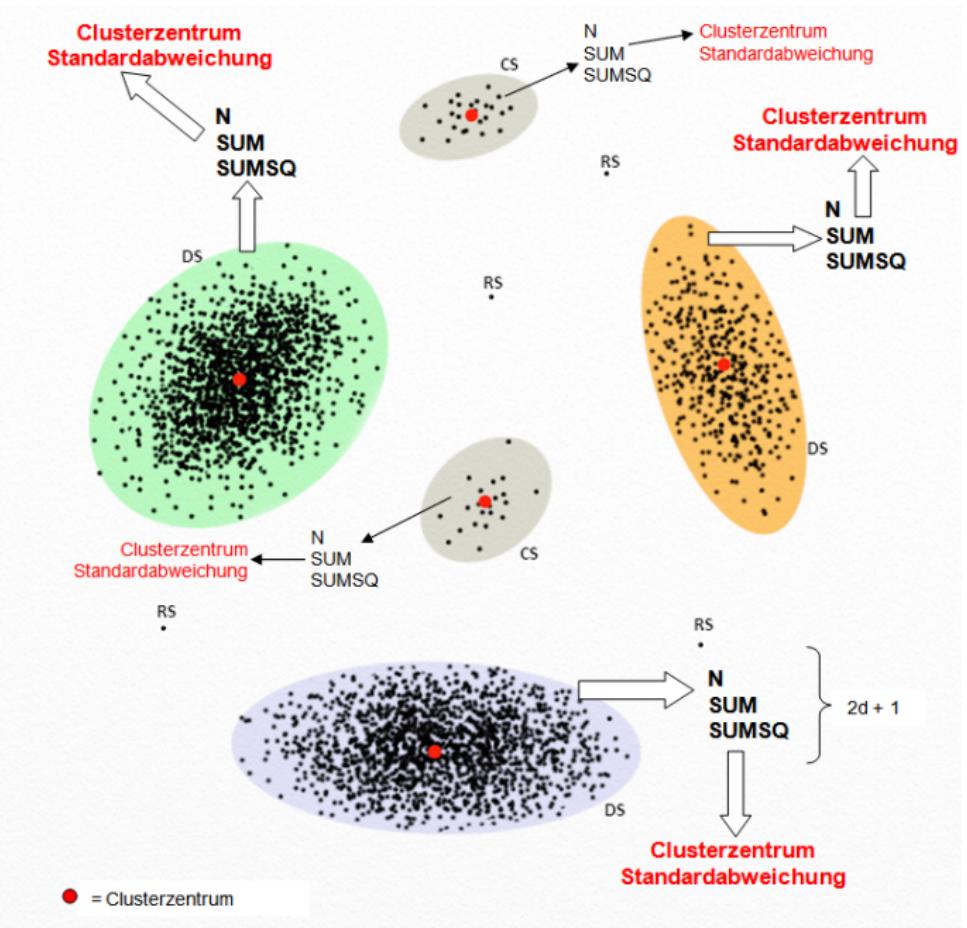












# Aktualisierung der Metadaten

Hinzufügen eines neuen Datenpunktes zu einem **DS** oder **CS**:

# Aktualisierung der Metadaten

Hinzufügen eines neuen Datenpunktes zu einem **DS** oder **CS**:

- Erhöhen der Anzahl der Datenpunkte  $\Rightarrow N + 1$

# Aktualisierung der Metadaten

Hinzufügen eines neuen Datenpunktes zu einem **DS** oder **CS**:

- Erhöhen der Anzahl der Datenpunkte  $\Rightarrow N + 1$
- Koordinaten des Punktes zum SUM-Vektor addieren

# Aktualisierung der Metadaten

Hinzufügen eines neuen Datenpunktes zu einem **DS** oder **CS**:

- Erhöhen der Anzahl der Datenpunkte  $\Rightarrow N + 1$
- Koordinaten des Punktes zum SUM-Vektor addieren
- Quadrate der Punktkoordinaten zum SUMSQ-Vektor addieren

# Aktualisierung der Metadaten

Hinzufügen eines neuen Datenpunktes zu einem **DS** oder **CS**:

- Erhöhen der Anzahl der Datenpunkte  $\Rightarrow N + 1$
- Koordinaten des Punktes zum SUM-Vektor addieren
- Quadrate der Punktkoordinaten zum SUMSQ-Vektor addieren

Hinzufügen eines neuen Datenpunktes zum **RS**:

# Aktualisierung der Metadaten

Hinzufügen eines neuen Datenpunktes zu einem **DS** oder **CS**:

- Erhöhen der Anzahl der Datenpunkte  $\Rightarrow N + 1$
- Koordinaten des Punktes zum SUM-Vektor addieren
- Quadrate der Punktkoordinaten zum SUMSQ-Vektor addieren

Hinzufügen eines neuen Datenpunktes zum **RS**:

$\Rightarrow$  der Datenpunkt selbst wird gespeichert

# Zuordnung der Punkte zu den Datensets

- 1 Punkt liegt nahe genug an einem Clusterzentrum (DS oder CS)  
⇒ Zuordnung zum DS bzw. CS

# Zuordnung der Punkte zu den Datensets

- ① Punkt liegt nahe genug an einem Clusterzentrum (DS oder CS)  
⇒ Zuordnung zum DS bzw. CS
- ② Punkt liegt nicht nahe genug an einem Clusterzentrum:

# Zuordnung der Punkte zu den Datensets

- ① Punkt liegt nahe genug an einem Clusterzentrum (DS oder CS)  
⇒ Zuordnung zum DS bzw. CS
- ② Punkt liegt nicht nahe genug an einem Clusterzentrum:
  - Punkt liegt nahe genug bei anderen Punkten des RS:

# Zuordnung der Punkte zu den Datensets

- ① Punkt liegt nahe genug an einem Clusterzentrum (DS oder CS)  
⇒ Zuordnung zum DS bzw. CS
- ② Punkt liegt nicht nahe genug an einem Clusterzentrum:
  - Punkt liegt nahe genug bei anderen Punkten des RS:  
⇒ zu neuem CS zusammenfassen und aus dem RS löschen

# Zuordnung der Punkte zu den Datensets

- ① Punkt liegt nahe genug an einem Clusterzentrum (DS oder CS)  
⇒ Zuordnung zum DS bzw. CS
- ② Punkt liegt nicht nahe genug an einem Clusterzentrum:
  - Punkt liegt nahe genug bei anderen Punkten des RS:  
⇒ zu neuem CS zusammenfassen und aus dem RS löschen  
⇒ ggf. neues CS mit bereits vorhandenen CS zusammenfassen

# Zuordnung der Punkte zu den Datensets

- ① Punkt liegt nahe genug an einem Clusterzentrum (DS oder CS)  
⇒ Zuordnung zum DS bzw. CS
- ② Punkt liegt nicht nahe genug an einem Clusterzentrum:
  - Punkt liegt nahe genug bei anderen Punkten des RS:  
⇒ zu neuem CS zusammenfassen und aus dem RS löschen  
⇒ ggf. neues CS mit bereits vorhandenen CS zusammenfassen
  - Datenpunkt liegt isoliert:  
⇒ Punkt im RS speichern

# Abschließende Zuordnung der Datensets

**Was passiert mit den CS und RS, nachdem alle Datenpunkte einem Set zugewiesen wurden?**

# Abschließende Zuordnung der Datensets

**Was passiert mit den CS und RS, nachdem alle Datenpunkte einem Set zugewiesen wurden?**

- alle CS und alle Punkte des RS dem nächsten DS-Clusterzentrum zuweisen

# Abschließende Zuordnung der Datensets

**Was passiert mit den CS und RS, nachdem alle Datenpunkte einem Set zugewiesen wurden?**

- alle CS und alle Punkte des RS dem nächsten DS-Clusterzentrum zuweisen
- Punkte des RS als Ausreißer behandeln

# Abschließende Zuordnung der Datensets

**Was passiert mit den CS und RS, nachdem alle Datenpunkte einem Set zugewiesen wurden?**

- alle CS und alle Punkte des RS dem nächsten DS-Clusterzentrum zuweisen
- Punkte des RS als Ausreißer behandeln
- CS als separate Minicluster behandeln

## Wann liegt ein Punkt nahe genug an einem Clusterzentrum?

① Jedes DS und CS enthält:

- Koordinaten des Clusterzentrums  $c = (c_1, \dots, c_d)$
- Standardabweichungen für jede Dimension  $s = (s_1, \dots, s_d)$

## Wann liegt ein Punkt nahe genug an einem Clusterzentrum?

① Jedes DS und CS enthält:

- Koordinaten des Clusterzentrums  $c = (c_1, \dots, c_d)$
- Standardabweichungen für jede Dimension  $s = (s_1, \dots, s_d)$

② Normalisierte (euklidische) Distanz  $z$  für jede Dimension  $i$ :

$$z_i = \frac{a_i - c_i}{s_i}$$

## Wann liegt ein Punkt nahe genug an einem Clusterzentrum?

① Jedes DS und CS enthält:

- Koordinaten des Clusterzentrums  $c = (c_1, \dots, c_d)$
- Standardabweichungen für jede Dimension  $s = (s_1, \dots, s_d)$

② Normalisierte (euklidische) Distanz  $z$  für jede Dimension  $i$ :

$$z_i = \frac{a_i - c_i}{s_i}$$

③ **Mahalanobis-Distanz:**

$$MD = \sqrt{\sum_{i=1}^d z_i^2}$$

## Wann liegt ein Punkt nahe genug an einem Clusterzentrum?

- ① Metadaten liefern Clusterzentren und Standardabweichungen
- ②  $a_i$  liegt in jeder Dimension genau die **einfache Standardabweichung** vom Clusterzentrum entfernt:

$$z_i = \frac{a_i - c_i}{s_i} = \frac{s_i}{s_i} = 1$$

- ③ **Mahalanobis-Distanz:**

$$MD = \sqrt{\sum_{i=1}^d z_i^2}$$

## Wann liegt ein Punkt nahe genug an einem Clusterzentrum?

- ① Metadaten liefern Clusterzentren und Standardabweichungen
- ②  $a_i$  liegt in jeder Dimension genau die **einfache Standardabweichung** vom Clusterzentrum entfernt:

$$z_i = \frac{a_i - c_i}{s_i} = \frac{s_i}{s_i} = 1$$

- ③ **Mahalanobis-Distanz:**

$$MD = \sqrt{\sum_{i=1}^d z_i^2} = \sqrt{\sum_{i=1}^d 1^2} = \sqrt{d}$$

## Wann liegt ein Punkt nahe genug an einem Clusterzentrum?

- ① Metadaten liefern Clusterzentren und Standardabweichungen
- ②  $a_i$  liegt in jeder Dimension genau die **einfache Standardabweichung** vom Clusterzentrum entfernt:

$$z_i = \frac{a_i - c_i}{s_i} = \frac{s_i}{s_i} = 1$$

- ③ **Mahalanobis-Distanz:**

$$MD = \sqrt{\sum_{i=1}^d z_i^2} = \sqrt{\sum_{i=1}^d 1^2} = \sqrt{d}$$

⇒ Grenzen für 2- bzw. 3-fache Standardabweichung:  $2\sqrt{d}$  bzw.  $3\sqrt{d}$

# Kritische Betrachtung des BFR-Algorithmus

## Vorteile

- für multivariate Daten geeignet

## Nachteile

# Kritische Betrachtung des BFR-Algorithmus

## Vorteile

- für multivariate Daten geeignet
- keine Iteration

## Nachteile

# Kritische Betrachtung des BFR-Algorithmus

## Vorteile

- für multivariate Daten geeignet
- keine Iteration
- generieren multipler Lösungen

## Nachteile

# Kritische Betrachtung des BFR-Algorithmus

## Vorteile

- für multivariate Daten geeignet
- keine Iteration
- generieren multipler Lösungen
- durch Metadaten geringerer Speicherplatzbedarf

## Nachteile

# Kritische Betrachtung des BFR-Algorithmus

## Vorteile

- für multivariate Daten geeignet
- keine Iteration
- generieren multipler Lösungen
- durch Metadaten geringerer Speicherplatzbedarf

## Nachteile

- einzelne Datenpunkte aus Metadaten nicht mehr extrahierbar

# Kritische Betrachtung des BFR-Algorithmus

## Vorteile

- für multivariate Daten geeignet
- keine Iteration
- generieren multipler Lösungen
- durch Metadaten geringerer Speicherplatzbedarf

## Nachteile

- einzelne Datenpunkte aus Metadaten nicht mehr extrahierbar
- für Hintergrundrauschen und Ausreißer anfällig

# Kritische Betrachtung des BFR-Algorithmus

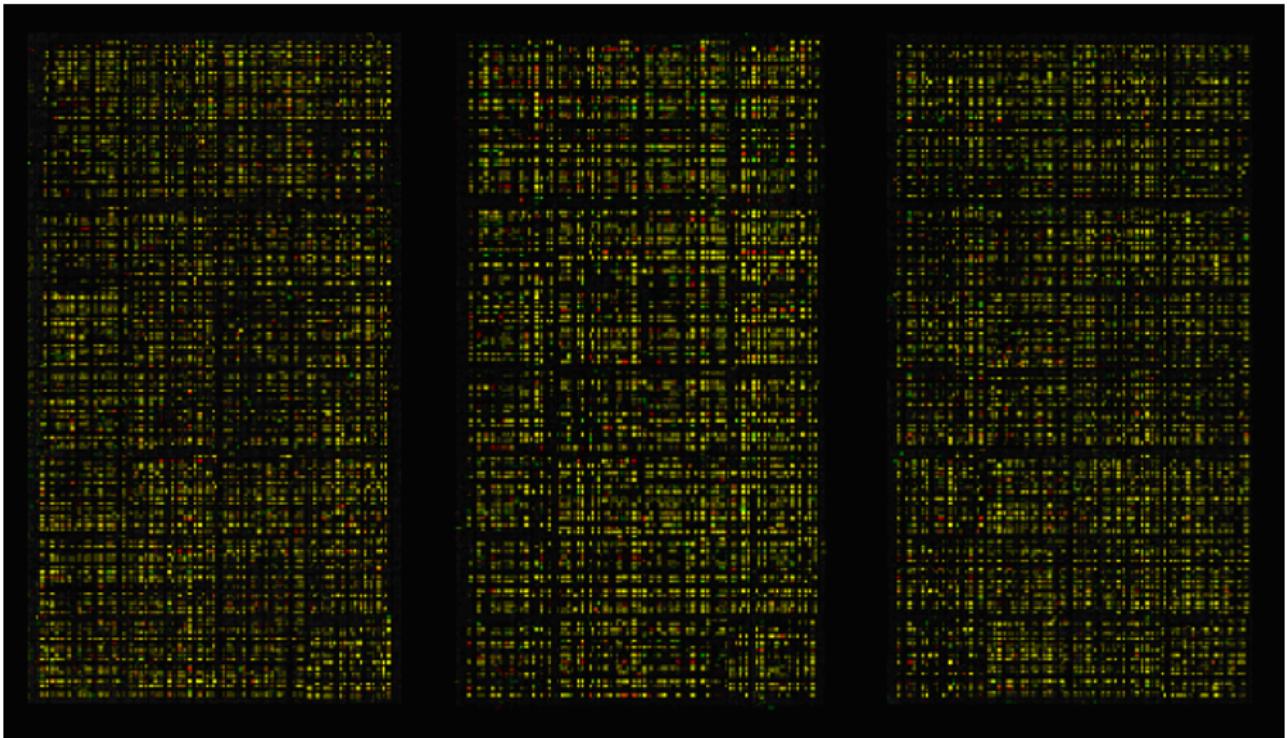
## Vorteile

- für multivariate Daten geeignet
- keine Iteration
- generieren multipler Lösungen
- durch Metadaten geringerer Speicherplatzbedarf

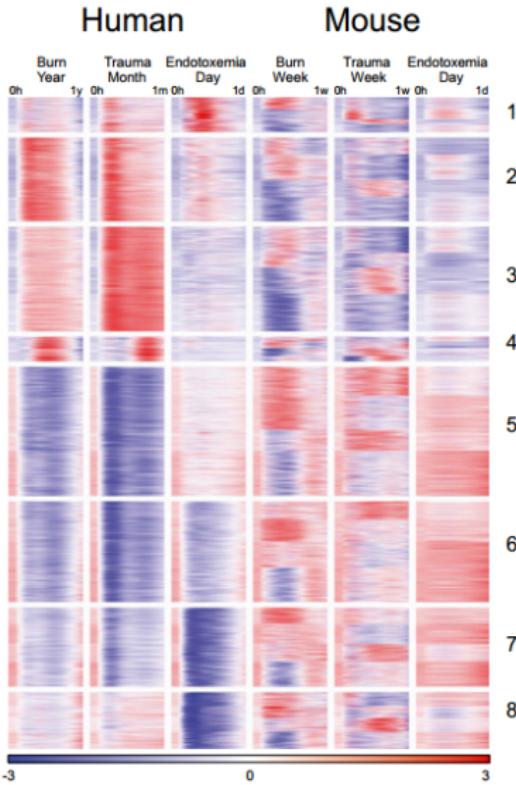
## Nachteile

- einzelne Datenpunkte aus Metadaten nicht mehr extrahierbar
- für Hintergrundrauschen und Ausreißer anfällig
- strenges Kriterium der Normalverteilung in jeder Dimension

# Anwendungsbeispiel: Genexpressionsanalysen



# Anwendungsbeispiel: Genexpressionsanalysen



Bildquelle: Seok et al. (2013)

# Bildquellen

- K. Backhaus, B. Erichson, W. Plinke, and R. Weiber. *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung (Springer-Lehrbuch) (German Edition)*. Springer, 2008. ISBN 9783540850441.
- A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.
- J. Seok, H. S. Warren, and Cuenca. Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences*, 110(9):3507–3512, 2013.
- D. Stotz. Der k-means algorithmus. Mentorisierte Arbeit in Fachdidaktik Mathematik, July 2016.