# Seminar Business Intelligence

# Data Preprocessing II

Sabine Queckbörner



## Übersicht

- Reduktion
  - Aggregation von Daten
  - Mengenreduktion
  - Numerosity Reduction
  - Diskretisierung von Datenwerten
- Kompression
  - Kompression multidimensionaler Daten
  - Zeichenkettenkompression

REDUKTION

## Aggregation von Daten

- **Reduktion**
  - Aggregation von Daten
  - Mengenreduktion
  - Numerosity Reduction
  - Diskretisierung von Datenwerten
- Kompression
  - Kompression multidimensionaler Daten
  - Zeichenkettenkompression

REDUKTION

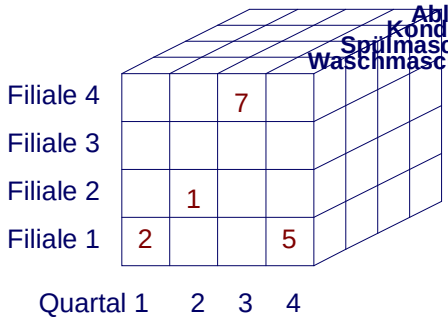
## Aggregation von Daten

- Data Cubes
  - verschiedene Abstraktionsebenen
  - Jede Abstraktionsebene reduziert die Größe der resultierenden Daten
  - Die höchste Abstraktionsebene und damit die größte Reduktion, wird mit der Aggregation aller Teilwürfel zu einem Gesamtwürfel erreicht.

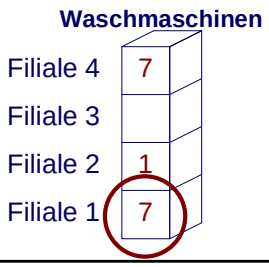
## REDUKTION

# Aggregation von Daten

- Data Cubes - Beispiel




**Wieviele Waschmaschinen wurden im ganzen Jahr in den Filialen verkauft?**

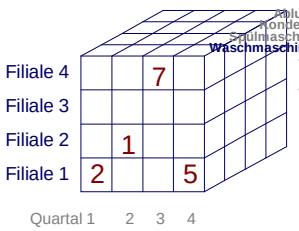




## REDUKTION

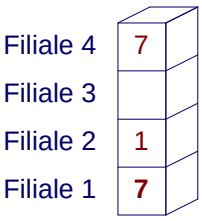
# Aggregation von Daten

- Data Cubes - Beispiel


Wieviele Waschmaschinen wurden im ganzen Jahr in den Filialen verkauft?

**Waschm.**




| Waschmaschinen | |
|---|---|
| Filiale 4 | 7 |
| Filiale 3 | - |
| Filiale 2 | 1 |
| Filiale 1 | 7 |

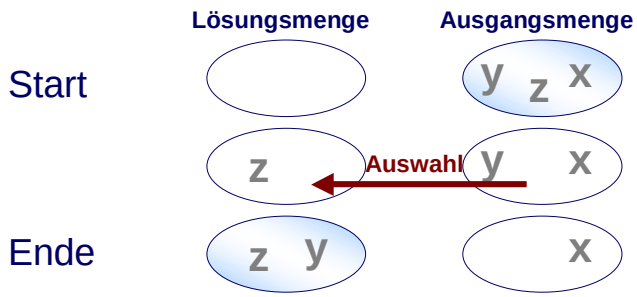
## Mengenreduktion

- **Reduktion**
  - Aggregation von Daten
  - Mengenreduktion
  - Numerosity Reduction
  - Diskretisierung von Datenwerten
- Kompression
  - Kompression multidimensionaler Daten
  - Zeichenkettenkompression



## Mengenreduktion

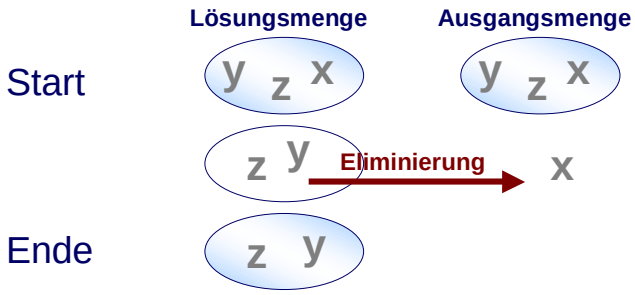
- Auswahlverfahren:
  - Schrittweise Vorwärtsauswahl



REDUKTION

## Mengenreduktion

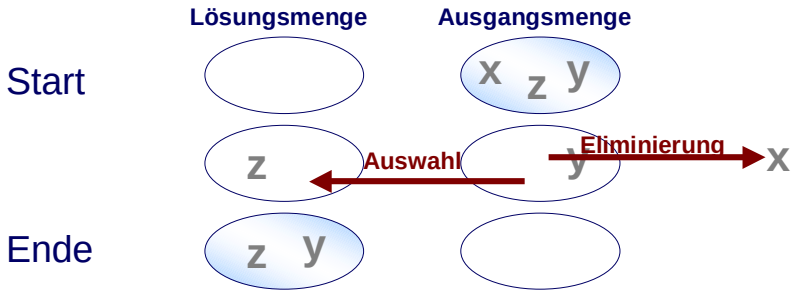
- Auswahlverfahren:
  - Schrittweise Vorwärtsauswahl
  - Schrittweise Rückwärtseliminierung




REDUKTION

## Mengenreduktion

- Auswahlverfahren:
  - Schrittweise Vorwärtsauswahl
  - Schrittweise Rückwärtseliminierung
  - Kombination aus beiden Verfahren



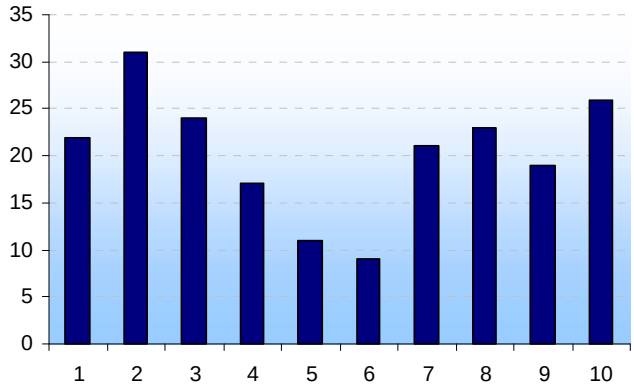
## Numerosity Reduction

- **Reduktion**
  - Aggregation von Daten
  - Mengenreduktion
  - Numerosity Reduction
  - Diskretisierung von Datenwerten
- Kompression
  - Kompression multidimensionaler Daten
  - Zeichenkettenkompression



## Numerosity Reduktion

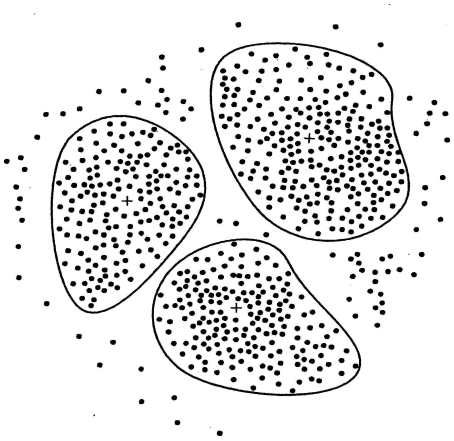
- **Histogramme**



REDUKTION

## Numerosity Reduktion

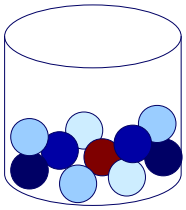
- Histogramme
- **Clustering**



REDUKTION

## Numerosity Reduktion

- Histogramme
- Clustering
- **Sampling**
  - Einfache zufällige Stichprobe ohne Ersetzung
  - Einfache zufällige Stichprobe mit Ersetzung
  - Cluster-Stichprobe
  - Schichtenweise Stichproben



REDUKTION

## Diskretisierung

- **Reduktion**
  - Aggregation von Daten
  - Mengenreduktion
  - Numerosity Reduction
  - Diskretisierung von Datenwerten
- Kompression
  - Kompression multidimensionaler Daten
  - Zeichenkettenkompression

REDUKTION

## Diskretisierung

- Numerische Daten
  - Binning
  - Histogramm - Analyse
  - Cluster - Analyse
  - Segmentierung durch natürliche Partitionierung
    - ð 3-4-5 Regel

REDUKTION

## Diskretisierung

- 3-4-5 Regel
  - 3, 6, 7 oder 9 unterschiedliche Werte ð Bereich in 3 Intervalle (2-3-2 für sieben) aufteilen
  - 2, 4 oder 8 unterschiedliche Werte ð Bereich in 4 gleichweite Intervalle aufteilen
  - 1, 5 oder 10 unterschiedliche Werte ð Bereich in 5 gleichweite Intervalle aufteilen

REDUKTION

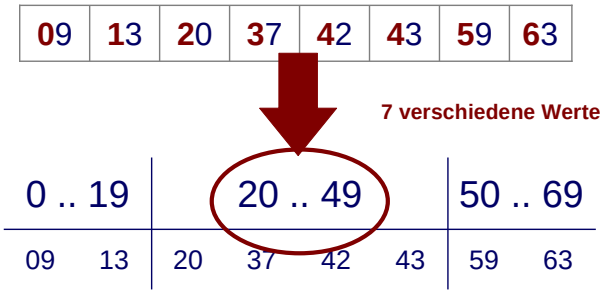
## Diskretisierung

- 3-4-5 Regel – Beispiel:

| 09 | 13 | 20 | 37 | 42 | 43 | 59 | 63 |
|---|---|---|---|---|---|---|---|

## Diskretisierung

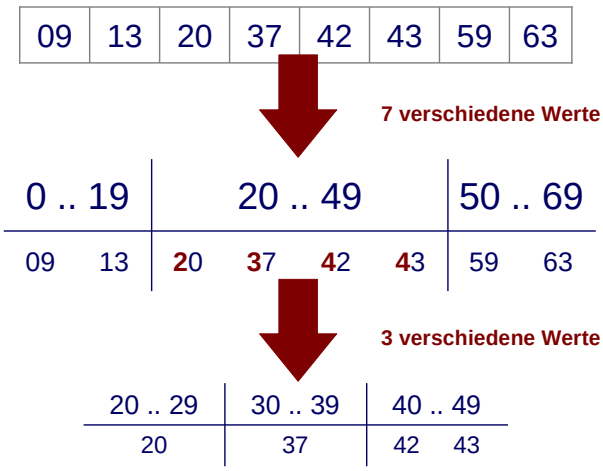
- 3-4-5 Regel – Beispiel:




REDUKTION

## Diskretisierung

- 3-4-5 Regel – Beispiel:



## KOMPRESSION
## Multidimensionale Daten

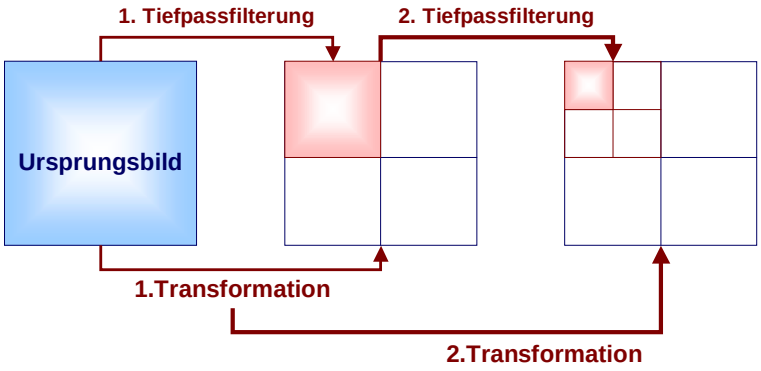
- Reduktion
  - Aggregation von Daten
  - Mengenreduktion
  - Numerosity Reduction
  - Diskretisierung von Datenwerten
- **Kompression**
  - Kompression multidimensionaler Daten
  - Zeichenkettenkompression

## KOMPRESSION
## Multidimensionale Daten

- Transformationskodierung – Prinzip:
  - Transformation
    - Daten *anders* darstellen
    - Zum Beispiel Fourier-Transformation, Cosinus-Transformation, Wavelet-Transformation
  - Quantisierung
    - Wertebereich der Bildpunkte einschränken
  - Kodierung
    - Binärkodierung
    - zum Beispiel durch Lauflängenkodierung

## Multidimensionale Daten

- Wavelet-Transformation - Prinzip:

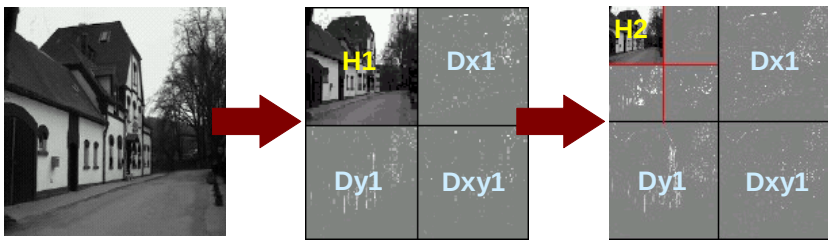



ð Immer kleiner werdende Hochpassanteile und ein einziger Tiefpassanteil

KOMPRESSION

## Multidimensionale Daten

- Wavelet-Transformation - Beispiel:



KOMPRESSION

## Zeichenkettenkompression

- Reduktion
  - Aggregation von Daten
  - Mengenreduktion
  - Numerosity Reduction
  - Diskretisierung von Datenwerten
- **Kompression**
  - Kompression multidimensionaler Daten
  - Zeichenkettenkompression

KOMPRESSION

## Zeichenkettenkompression

- Dictionary-basierte Algorithmen
- Statistische Kodierer
- Borrows-Wheeler-Transformation

KOMPRESSION

## Zeichenkettenkompression

- Dictionary-basierte Algorithmen
- Statistische Kodierer
- **Borrows-Wheeler-Transformation**
  - Zeichenkette umsortieren
  - kodieren

KOMPRESSION

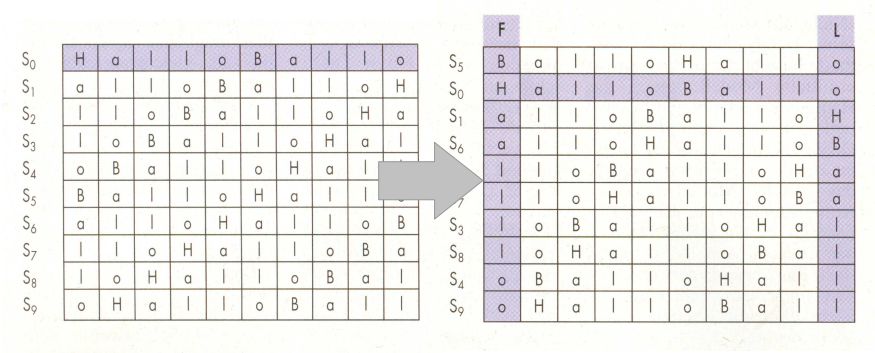
## Zeichenkettenkompression

- Dictionary-basierte Algorithmen
- Statistische Kodierer
- **Borrows-Wheeler-Transformation**
  - Zeichenkette umsortieren
    - quadratische Matrix erstellen
    - Zeilen der Matrix alphabetisch sortieren

    ð Ausgabe:
      - letzte Spalte
      - Position der Ausgangszeichenkette in sortierter Matrix
  - kodieren

# KOMPRESSION
## Zeichenkettenkompression

- Borrows-Wheeler-Transformation – Beispiel:

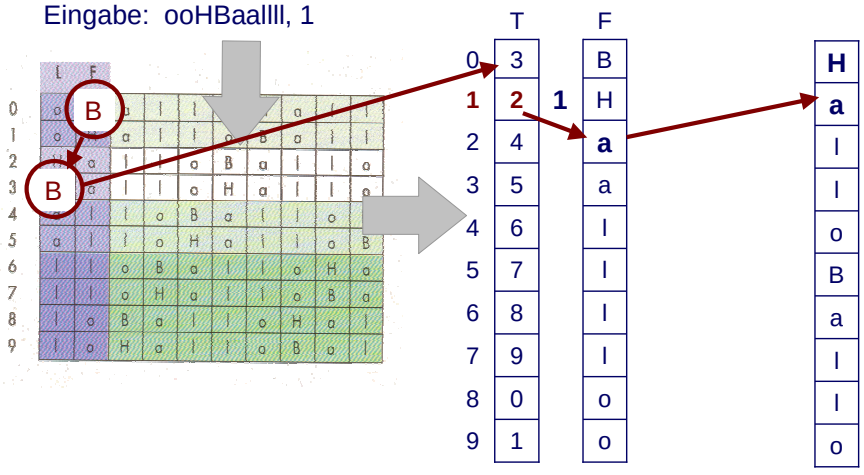



Ausgabe: ooHBaallll, 1

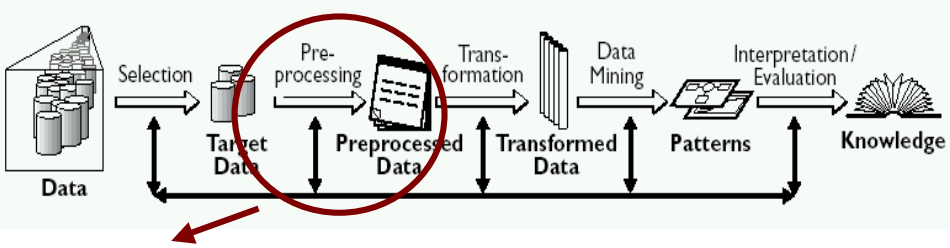
# KOMPRESSION
## Zeichenkettenkompression

- Borrows-Wheeler-Transformation – Beispiel:

Eingabe: ooHBaallll, 1



## Zusammenfassung




- Reduktion
  - Aggregation von Daten, Mengenreduktion, Numerosity Reduction, Diskretisierung
- Kompression
  - Multidimensionale Daten und Zeichenketten

## Vielen Dank