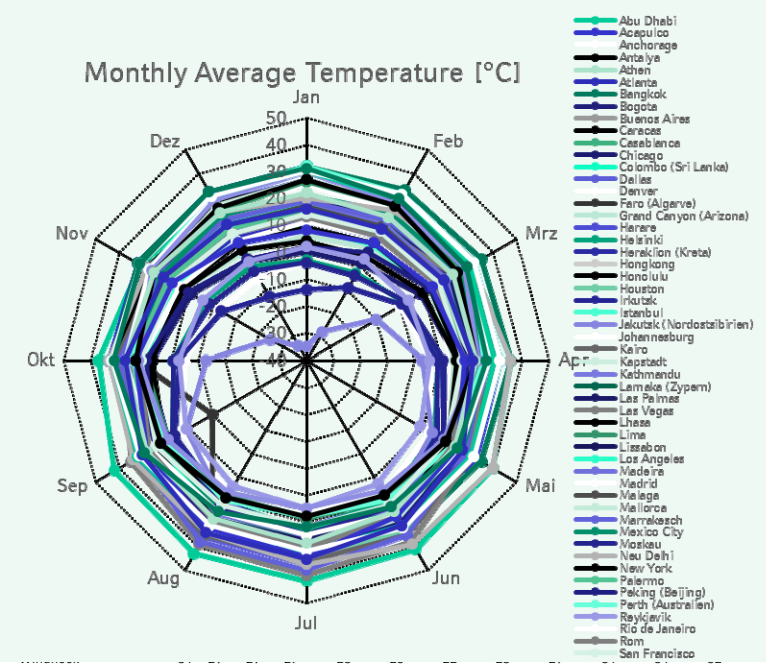
Daten Visualisierung

- Muster in großen Daten sind schwer anhand von tabellarischen Repräsentationen zu verstehen
- Transformation der Daten in eine visuell verständliche Repräsentation
  („a picture is worth a thousand words“)
- Kombination versch. Fähigkeiten
  - *Computers* are good in number crunching (and data visualization by means of computer graphics)
  - *Humans* are good in visual pattern recognition

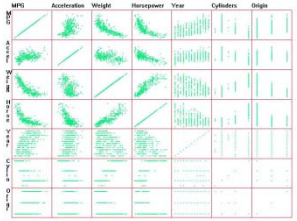
Monthly average temperature [°C]

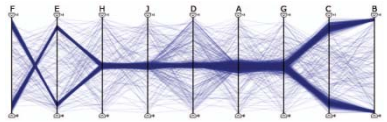
Städte Ø	Jan	Feb	Mrz	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
Abu Dhabi	25	27	31	36	40	41	42	43	42	37	31	27
Acapulco	32	31	32	32	33	33	33	33	33	33	32	32
Anchorage	-4	-2	0	6	13	17	18	17	13	5	-3	-5
Antalya	15	16	19	22	27	32	35	36	32	27	21	17
Athen	13	14	17	20	26	30	34	34	29	24	18	14
Atlanta	11	13	18	23	26	30	31	31	28	23	17	12
Bangkok	32	33	35	36	35	34	33	33	33	32	32	32
Bogota	20	19	19	19	19	18	18	18	19	19	19	20
Buenos Aires	30	28	26	23	19	16	15	17	19	21	26	29
Caracas	30	28	30	30	31	32	32	32	33	32	31	30
Casablanca	18	18	20	21	22	25	26	27	26	24	21	19
Chicago	0	1	9	16	21	26	29	28	24	17	9	2
Colombo (Sri Lanka)	31	31	32	32	32	31	31	31	31	31	31	31
Dallas	13	16	21	25	29	33	36	36	32	26	19	14
Denver	7	8	14	14	21	28	32	30	25	18	12	6
Faro (Algarve)	16	16	19	21	23	27	29	29	26	23	19	17
Grand Canyon (Arizona)	6	8	13	15	21	27	29	27	25	18	12	6
Harare	27	26	27	26	24	21	22	24	28	29	28	27
Helsinki	-3	-3	2	9	15	20	23	21	17	9	3	0



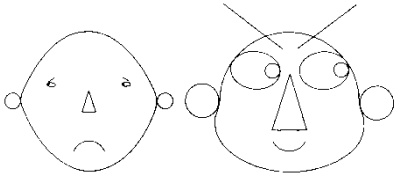
Daten Visualisierung: Techniken

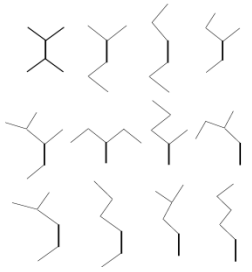
- Geometrische Techniken
  - Idea: Visualization of geometric transformations and projections of the data
  - Examples: Scatterplots, Parallel Coordinates



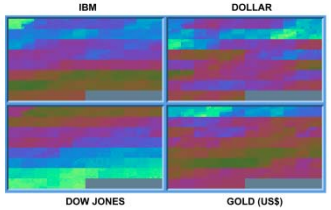


- Icon-basierte Techniken
  - Idee: Visualisierung durch Icons
  - Beispiele: Chernoff Faces, Stick Figures





- Pixel-orientierte Techniken
  - Idee: Visualize each attribute value of each data object by one colored pixel
  - Beispiele: Recursive Muster



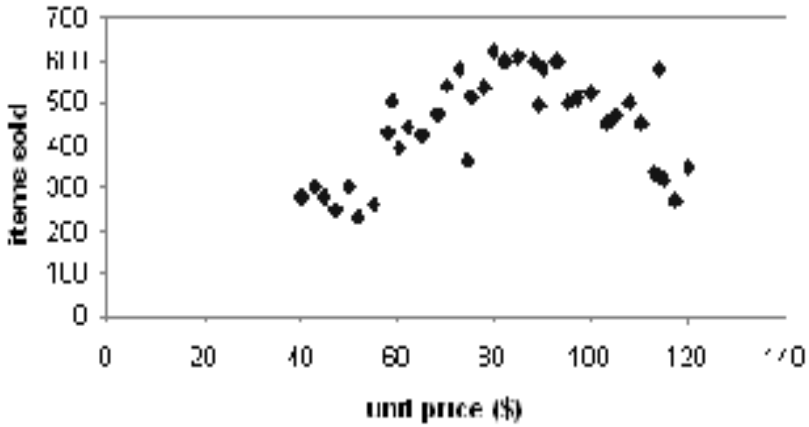
- Andere Techniken:
  - Hierarchical Techniques, Graph-based Techniques, Hybrid-Techniques, ...

Folie aus: Keim, Visual Techniques for Exploring Databases, Tutorial Slides, KDD 1997.

Beispiel: Scatter plot & Loess Curve

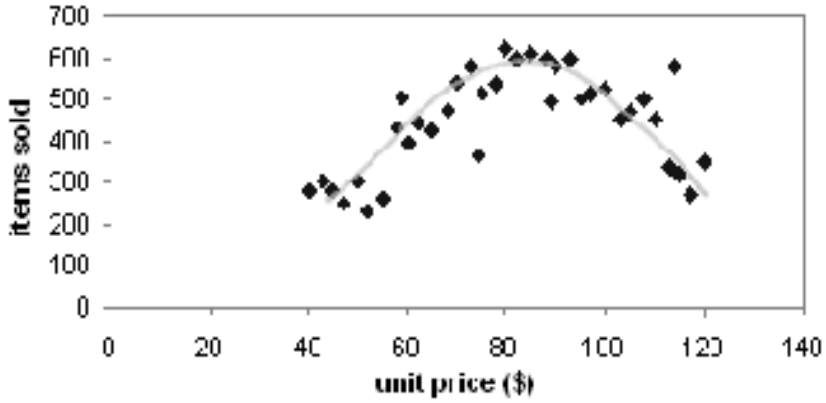
Scatter plot

- Erster Eindruck von 2D Daten, um Datenverteilung, Cluster, Outlier, etc. zu erkennen



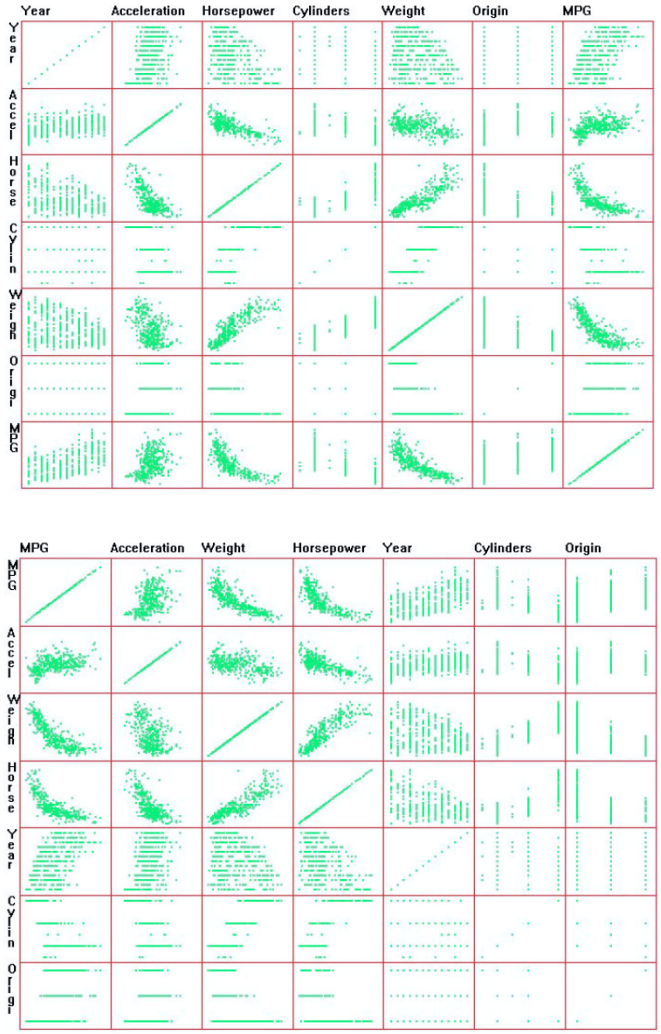
Loess Curve (local regression)

- Fittet eine Kurve auf einen Scatter Plot um die Abhängigkeiten (lokal) besser darzustellen
- Loess Curve ist abhängig von zwei Parametern:
  - Smoothing Parameter
  - Grad der Polynome, die durch Regression "gefittet" werden



Scatterplot Matrix

- Matrix von Scatterplots (x-y Diagramme) eines d-dimensionalen Datensatzes
  - Paarweise Koorelationen können erkannt warden
  - Keine komplexen Zusammenhänge
  - Sortierung der Dimensionen ist ggfls. Wichtig um unterschiedliche Zusammenhänge zu erkennen

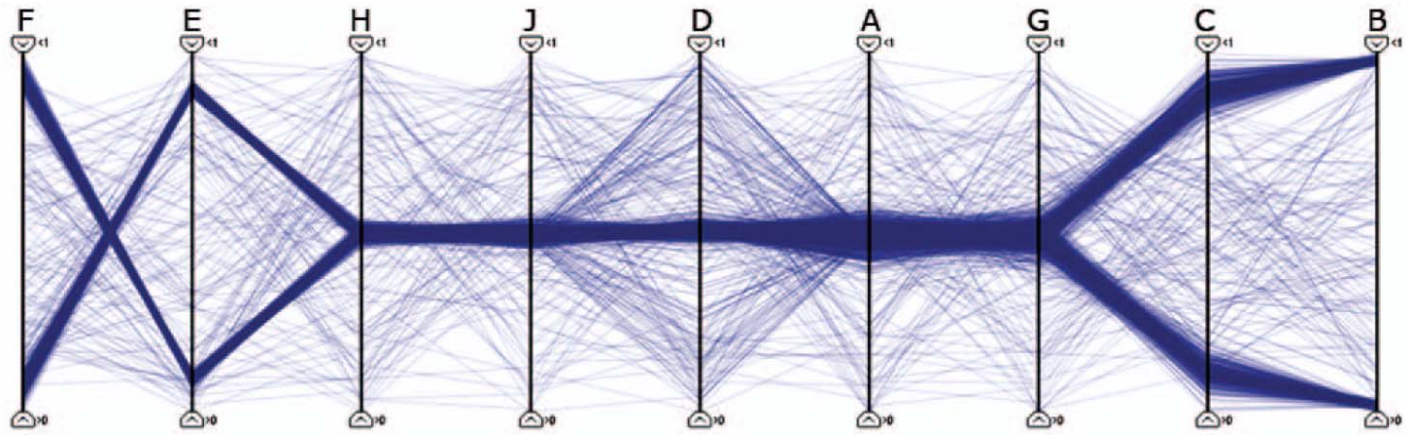


Figures from Peng et al., Clutter Reduction in Multi-Dimensional Data Visualizazion Using Dimension Reordering, IEEE Symp. on Inf. Vis., 2004.

Parallel Coordinates

[Ins 85] Inselberg, A.: The Plane with Parallel Coordinates, Special Issue on Computational Geometry. The Visual Computer, Vol. 1, pp. 69-97, 1985.

- Visualisierung eines d-dimensionalen Datensatzes mit d parallelen Achsen
- Jede Achse wird auf ihren min-max Wertebereich skaliert
- Jedes Datenobjekt ist eine Polygon-Line, die jede Achse an dem Punkt schneidet, die den Wert des Objekts in dieser Dimension darstellt



Slide credit: Keim, Visual Techniques for Exploring Databases, Tutorial Slides, KDD 1997.

Figure from Bertini et al., Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization, Trans. on Vis. and Comp. Graph., 2011.

Parallel Coordinates

- Achtung: Sortierung der Dimensionen wichtig!!!
- Relevanz/Interessantheit einer Sortierung kann quantitativ gemessen werden

- Beispiel:
  - Die erste Sortierung ist gut geeignet für die Visualisierung von Clustern
  - Die zweite Sortierung ist gut geeignet für die Visualisierung von Korrelationen zwischen den Merkmalen

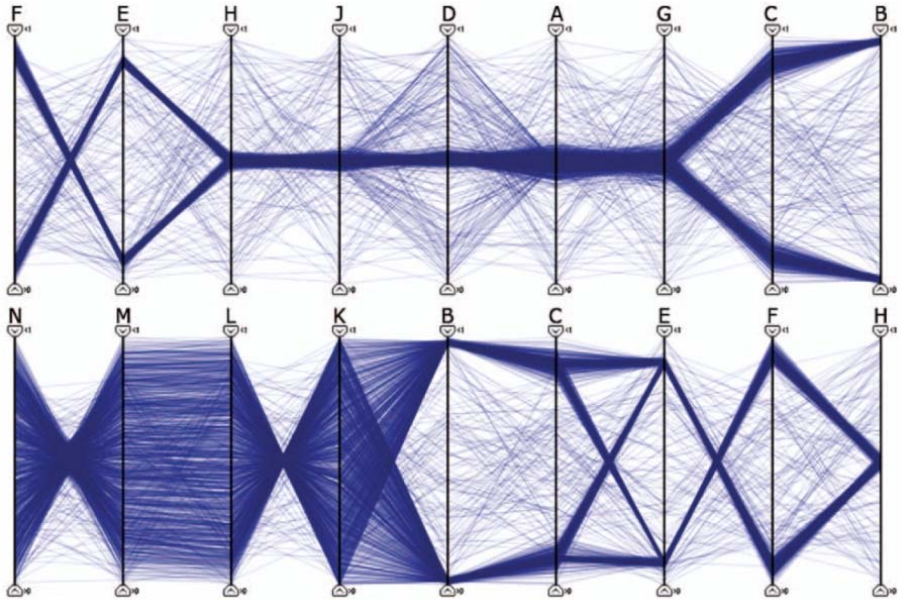
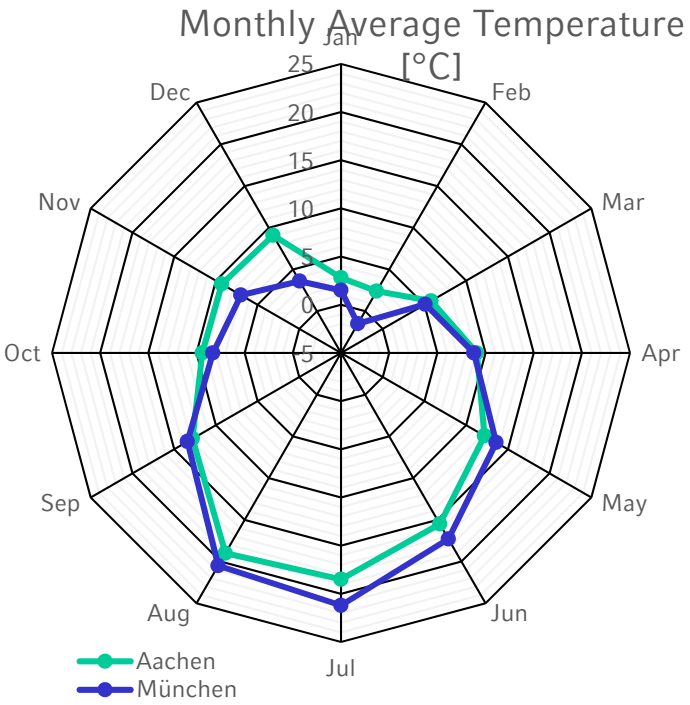
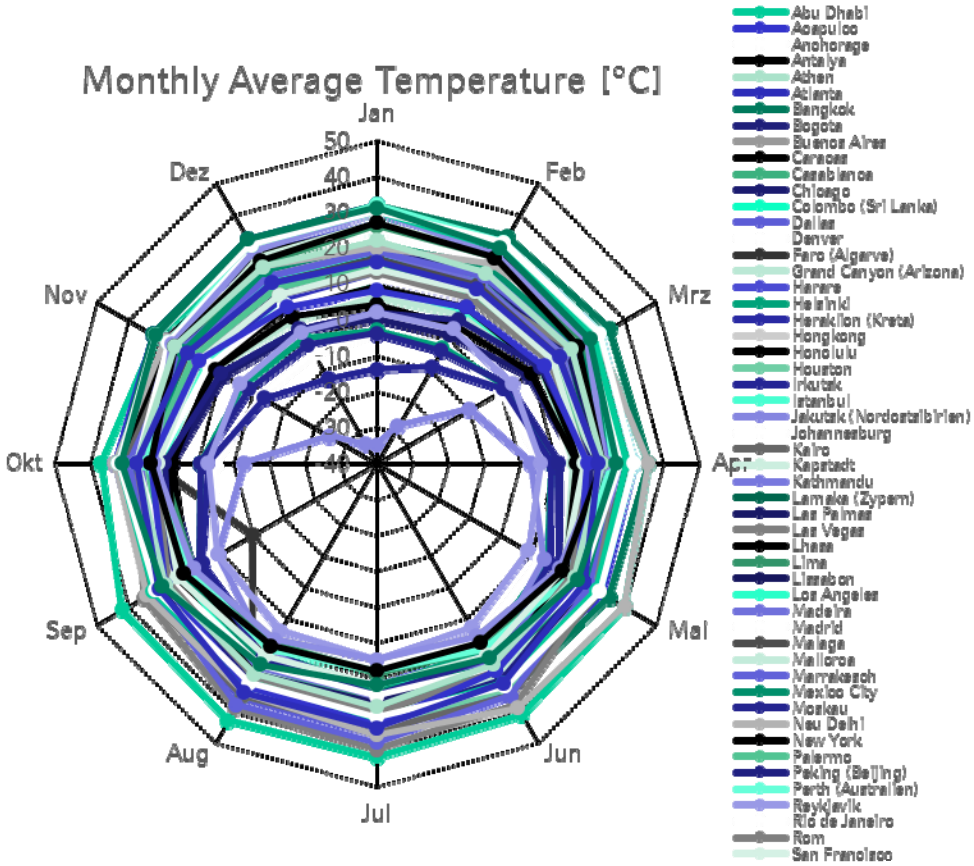


Figure from Bertini et al., Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization, Trans. on Vis. and Comp. Graph., 2011.

Spiderweb Modell

- Ähnlich wie bei Parallel Coordinates aber Objekte werden als Polylinien in einem „Spinnennetz“ angeordnet (Mitte = Ursprung aller Dimensionen)
- Nur für wenige Datenpunkte geeignet (wird schnell unübersichtlich)





Überblick

- Daten Repräsentation
  - Datentypen
  - Vergleich von Datenobjekten, Ähnlichkeit
  - Daten Visualisierung

- Data Warehousing
  - Daten Reduktion (Data Reduction)
  - Aggregation/Generalisierung

Data Reduction

- Motivation
  - DR hilft, Muster besser zu verstehen
    - Rohdaten (typ. Tabellen) sind schwer zu überblicken/verstehen
    - Visualisierungstechniken skalieren meist nur bis zu einer 4-stelligen Menge von Daten
    - DR kann helfen, (einfache) Muster zu erkennen (ohne DM Algorithmen)

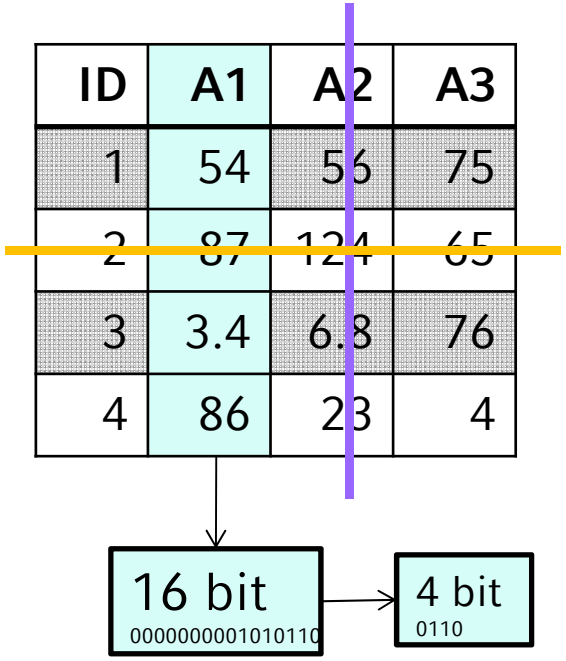
    => Die Idee von Business Intelligence (BI) basiert im Wesentlichen auf dieser Idee

  - (Berechnungs-) Komplexität
    - Big Data => hohe Laufzeiten von Data Mining Algorithmen
    - Reduzierte Daten führen meist auch zu reduzierten Laufzeiten

    (Nutzen ist allerdings nur dann gegeben, wenn die Algorithmen auf den reduzierten Daten (annähernd) die selben analytischen Resultate produzieren)

- 2 Arten der Daten Reduktion
  - Generalisieren von Daten (Abstraktion auf ein höheres Abstraktionsniveau)
  - Aggregation von Daten (durch grundlegende deskriptive Statistik)

Data Reduction Strategien



Numerosity reduction

Reduziere die Anzahl der Objekte

Dimensionality reduction

Reduziere die Anzahl der Attribute

Quantization/Discretization

Reduziere die Anzahl der Werte pro Attributs-Domäne

ID	A1	A3
1	L	75
3	XS	76
4	XL	4

Beispiele für Reduzierungs-Techniken

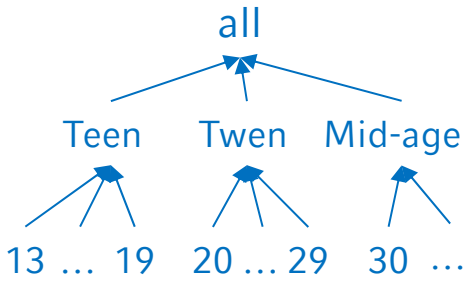
- Numerosity reduction
  - Sampling (=> Verlust von Daten)
  - Aggregation (abh. vom verwendeten Aggregations-Model und dessen Parameter, z.B. Center / Spread)

- Dimensionality reduction
  - Lineare Methoden: feature subselection; principal components analysis (PCA); random projections; Fourier transform; wavelet transform
  - Nicht-lineare Methoden: Multidimensional scaling (force model)

- Quantization
  - Binning, d.h. Transformation mit einem Histogramm (versch. Typen gebräuchlich)
  - Generalisierung entlang Konzept-Hierarchien (OLAP; „attribute-oriented induction“)

Aggregierte Repräsentationen

- Zusammenhang zwischen Quantisierung, Dimensions-Reduktion und Generalisierung
  - Quantisierung ist ein Spezialfall von Generalisierung
    - Beisp.: group *age* (7 bits) to *age_range* (4 bits)
  - Dimensions-Reduktion ist eine degenerierte Quantisierung
    - Dropping *age* reduces 7 bits to zero bits
    - Corresponds to (trivial) generalization of *age* to „all“ = „any age“ = no information



- Generalisierung erzeugt Duplikate
  - Eliminiere Duplikate aber speichere die urspr. Anzahl durch zus. Zählattribute
  - Aggregation ist „numerosity reduction“ (=> weniger Objekte)

Name	Age	Major
Ann	27	CS
Bob	26	CS
Eve	19	CS

Generali-zation

Name	Age	Major
(any)	Twen	CS
(any)	Twen	CS
(any)	Teen	CS

Aggre-gation

Age	Major	Count
Twen	CS	2
Teen	CS	1

Grundlegende Aggregationen

- Einfache Kennzahlen aus der deskriptiven Statistik
  - Central tendency: Wo ist das Zentrum der Daten?
    - Beispiele: Mittelwert, Median, Mode, …
  - Variation, spread: Wie stark ist die Abweichung vom Zentrum?
    - Beispiele: Varianz / Standard-Abweichung, min-max-Bereich, …

- Beispiele
  - Das Alter der Studenten ist um 20+
  - Schuhgröße ist zentriert um 40
  - Durchschnittsverdienst liegt im Bereich 1000 bis 5000

Distributive Aggregationsfunktionen

Distributiv | Algebraisch | Holistisch

- *Distributiv*
  - Das Resultat der Funktion angewendet auf *n* aggregierte Werte ist identisch mit dem angewendet auf alle Daten (ohne vorherige Aggregierung)

- Beispiele:

$$count(D_1 \cup D_2) = count(D_1) + count(D_2)$$
$$sum(D_1 \cup D_2) = sum(D_1) + sum(D_2)$$
$$min(D_1 \cup D_2) = \min(\min(D_1), \min(D_2))$$
$$max(D_1 \cup D_2) = \max(\max(D_1), \max(D_2))$$

Algebraische Aggregationsfunktionen

Distributiv | Algebraisch | Holistisch

- *Algebraisch*
  - Basiert auf einer algebraischen Funktion mit $M \in \aleph$ Argumenten, die jeweils durch die Anwendung einer distributiven Aggregatsfunktion berechnet wurden.
  - Beispiele: $avg() = sum() \,/\, count()$; $standard\_deviation()$

- Algebraische Funktionen sind meist nicht distributive, z.B.:

$$avg(D_1 \cup D_2) = \frac{sum(D_1 \cup D_2)}{count(D_1 \cup D_2)} = \frac{sum(D_1) + sum(D_2)}{count(D_1) + count(D_2)}$$

$$\neq avg\big(avg(D_1), avg(D_2)\big)$$

Holistische Aggregationsfunktionen

Distributiv | Algebraisch | Holistisch

- *Holistisch*
  - Hier gibt es keine Begrenzung des Speicherplatzes, der notwendig ist, um (Unter-) Aggregate zu bestimmen/beschreiben
- Examples:
  - $median$: value in the middle of a sorted series of values (= 50% quantile)
  - $median(D_1 \cup D_2) \neq simple\_function(median(D_1), median(D_2))$
  - $mode$: value that appears most often in a set of values
  - $rank$: k-smallest / k-largest value (cf. quantiles, percentiles)

Central Tendency Maße

- *Mean*
  - (gewichteter) aritmetischer Mittelwert

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Algebraisches Maß, nur für nummerische Daten (sum, scalar multiplication)

- *Mid-range*
  - Mitte zwischen dem größten und dem kleinsten Wert in einem Datensatz:

$$(\max + \min) / 2$$

  - Nur für nummerische Daten geeignet

→ Kategorische Daten?

Central Tendency Maße

- *Median*
  - Der mittlere Wert einer ungeraden Anzahl von Werten
  - Bei gerader Anzahl: Durchschnitt der beiden mittleren Werte (nummerische Daten), oder eine der beiden mittleren Werte (nicht-nummerische Daten)
  - Beispiele
    - never, never, never, rarely, rarely, often, usually, usually, always
    - tiny, small, big, big, big, big, big, big, huge, huge
    - tiny, tiny, small, medium, big, big, large, huge
  - Holistisches Maß, nur für ordinale Daten geeignet (Ordnung)

→ Was, wenn es keine Ordnung gibt?

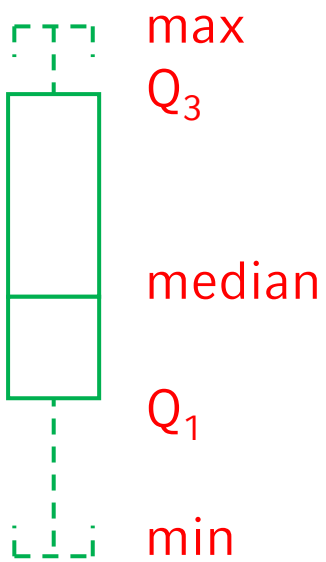
Central Tendency Maße

- *Mode*
  - Der Wert, der am häufigsten in den Daten vorkommt
  - Beispiel: blue, red, blue, yellow, green, blue, red
  - Unimodal, bimodal, trimodal, …: dann gibt es 1, 2, 3, … Modes, cf. Mixture Models (später)
  - Kommen alle Werte nur genau einmal vor, gibt es keinen Mode
  - Empirische Formel für unimodal Häufgkeitsverteilungen, die "moderately skewed" sind:

$$mean - mode \approx 3 \cdot (mean - median)$$

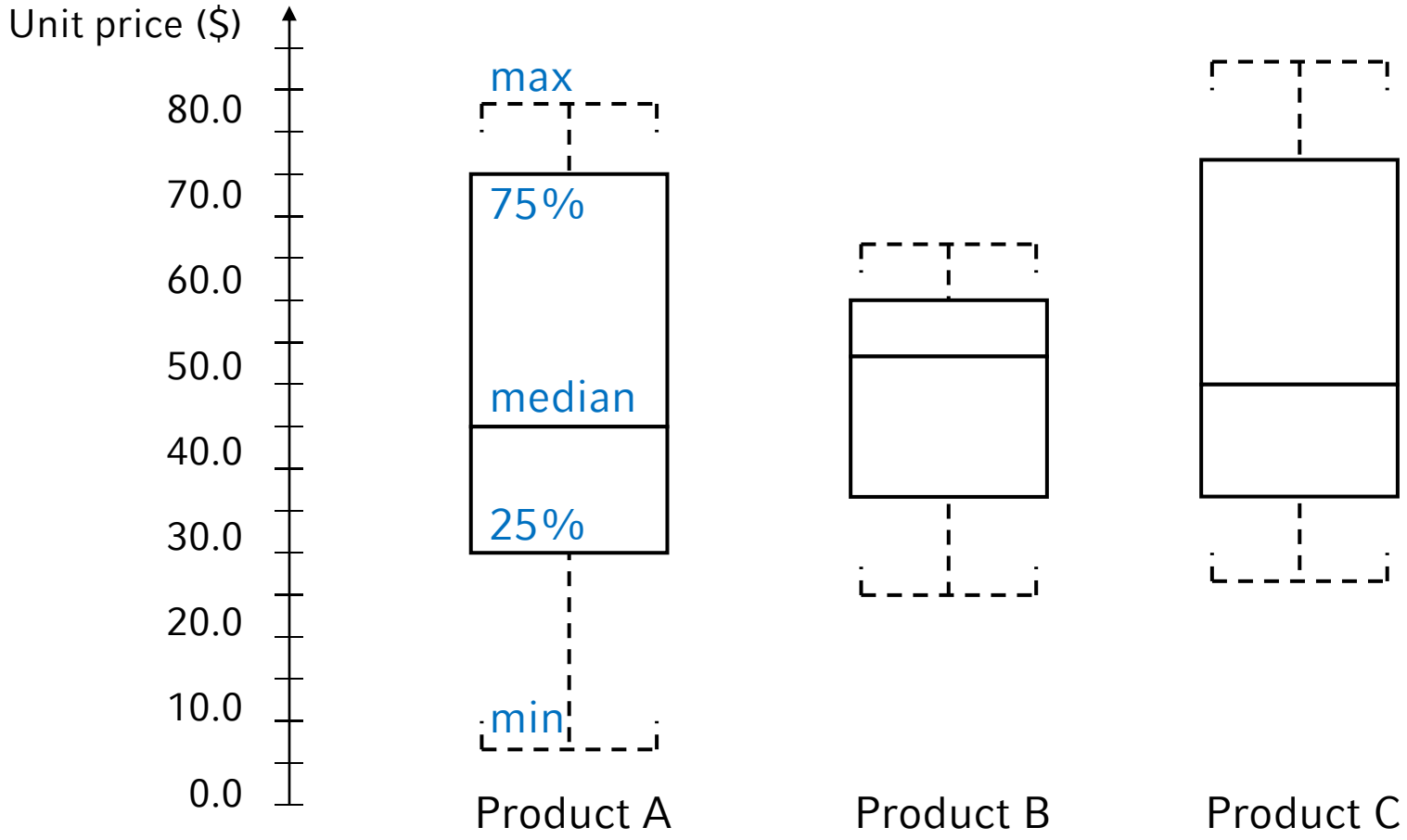
Boxplots

- 5 Werte einer Verteilung auf einen Blick
  - Minimum, Q1, Median, Q3, Maximum (Q = Quartil)
  - Repräsentiert 0%, 25%, 50%, 75%, 100%-Quantil der Daten
  - aka “25-percentile”, etc.

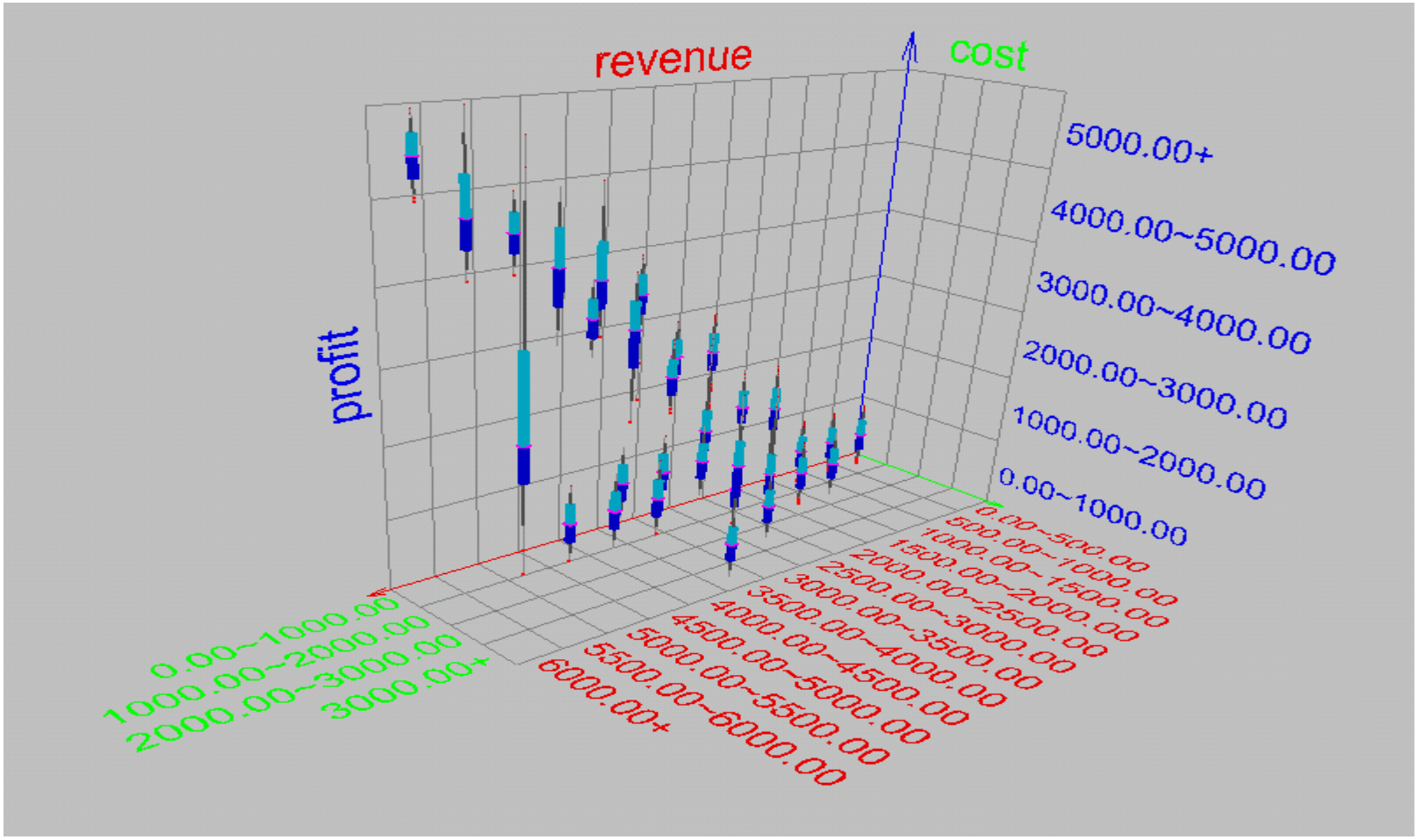


- Boxplot
  - Repräsentiert einen Datensatz durch eine Box
  - Grenzen der Box: Q1 und Q3
  - Höhe der Box: inter-quartile range $IQR = Q_3 - Q_1$
  - Median: Line innerhalb der Box
  - Whiskers: zwei Linien außerhalb der Box bis Minimum und Maximum
  - Outliers: typw. Werte, die 1.5 x IQR unter Q_1 oder über Q_3

Boxplots: Beispiele



Multidimensionale Boxplot Analyse



Generalisierung

- Generalisierung: Aggregation einer Menge von Daten (z.B. mit den oben beschrieben Maßen)
- Problem: Welche Partitionen der Daten sollen aggregiert werden?
- Alle Daten
  - Mittelwert und Varianz über alles → zu grob („overgeneralized“)
- Unterschiedliche Techniken um Daten für die Aggregation zu gruppieren
  - Binning – Histogramme, basierend auf Wertebereiche
  - Generalisierungshierarchie – Abstraktion basierend auf Konzept

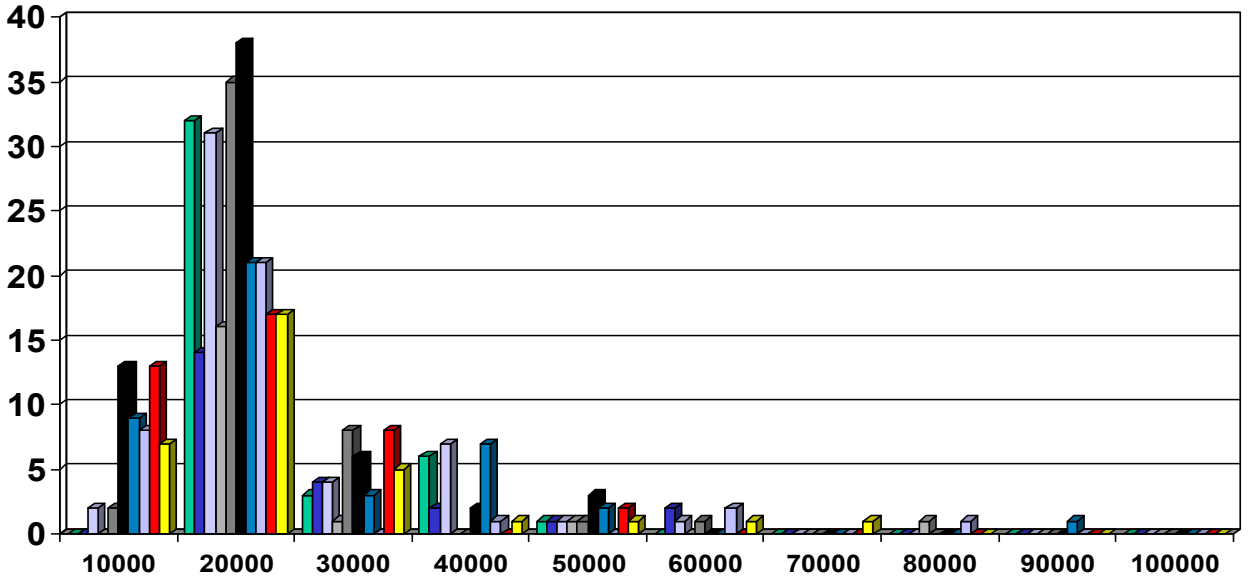
  } In BI-Lösungen typw. implementiert

  - Clustering (siehe später) – basierend auf Objekt-Ähnlihckeit

  } In BI-Lösungen typw. NICHT implementiert

Binning Techniken: Histogramme

- Histogramme approximieren die Datenverteilung mittels Binning
- Verteile die Daten auf die Bins und verwalte aggregierte Werte (sum, average, median) für jedes Bin
- Sehr populär
- Verwand mit dem Problem der Quantisierung



Equi-width Histogramme

- Wertebereich wird in N Intervalle gleicher Größe (Uniform Grid)
- Breite der Intervalle $W = (max - min)/N$

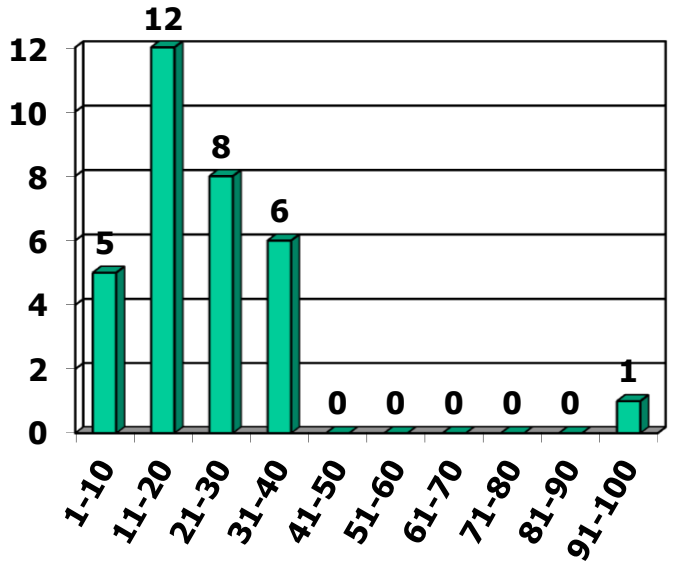
\+ einfach

– Outlier dominieren

– Probleme bei “skewed data”

- Beispiele (sortierte Daten, 10 bins):

  5, 7, 8, 8, 9, 11, 13, 13, 14, 14,
  14, 15, 17, 17, 17, 18, 19, 23, 24,
  25, 26, 26, 26, 27, 28, 32, 34, 36,
  37, 38, 39, 97



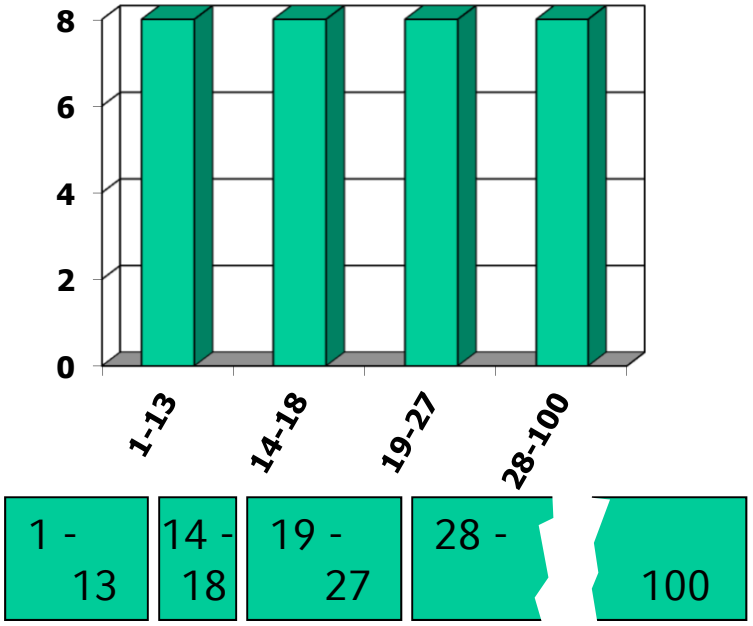
=> Was passiert beim Einfügen von 1023

Equi-height Histogramme

- Alle N Intervalle enthalten (ungefähr) gleich viele Daten (*quantile-based approach*)

+ Skaliert auch für dehr große Mengen

– Wenn ein Wert sehr häufig vorkommt, müssen Intervalle vereinigt werden



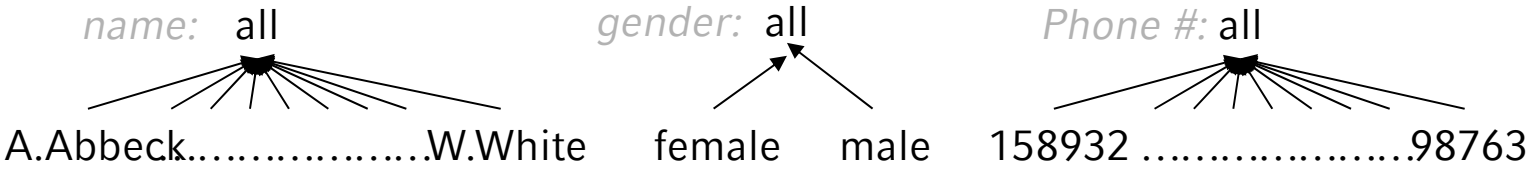
- Beispiel von oben (4 bins):

  5, 7, 8, 8, 9, 11, 13, 13, 14, 14,
  14, 15, 17, 17, 17, 18, 19, 23, 24,
  25, 26, 26, 26, 27, 28, 32, 34,
  36, 37, 37, 38, 97

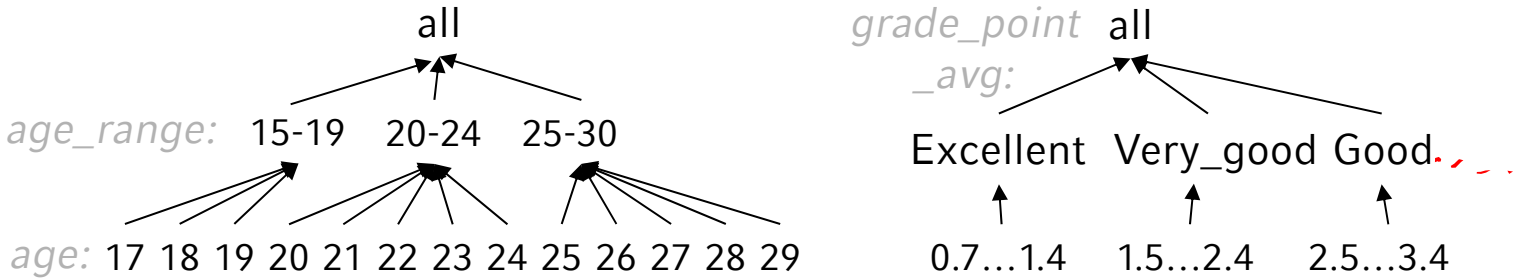
- Ein Wert? Median = 50%-Quantil
  - Ist robuster gegen Outlier (cf. Wert 1023 einfügen)
  - Vier Bins sind sehr ähnlich wie Boxplots

Concept Hierarchies: *Beispiele*

no (real) hierarchies



set grouping hierarchies



schema hierarchies