UNIVERSITÄT
LEIPZIG

# Informationsvisualisierung

Sommersemester 2025

Dirk Zeckzer

Institut für Informatik

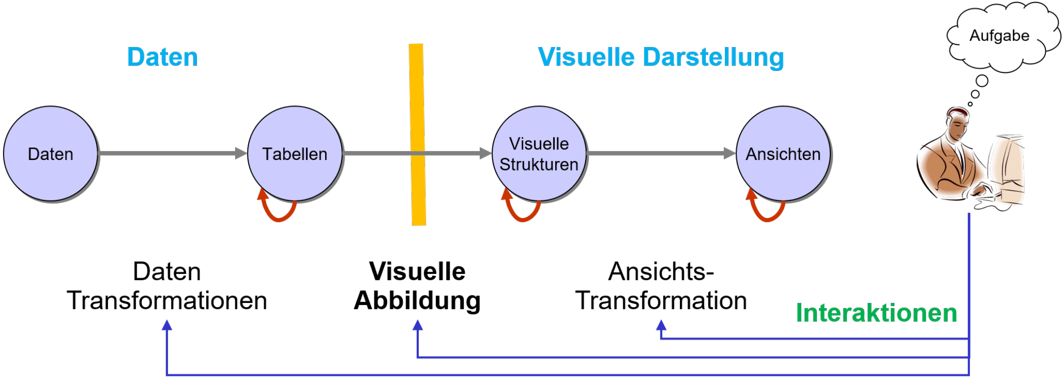
# Teil V

# Darstellung von mehrdimensionalen Daten

# Übersicht

5. Darstellung von mehrdimensionalen Daten
   - 5.1 Visuelle Abbildung
   - 5.2 Graphische Elemente
   - 5.3 Uni-Variate Daten
   - 5.4 Bi-Variate Daten
   - 5.5 Tri-Variate Daten
   - 5.6 Multi-Variate Daten

# Visuelle Abbildung



# Graphische Elemente

- 2D
  - Punkte (Pixel)
  - Ebene geometrische Figuren
    - Kreise, Ellipsen
    - Quadrate, Rechtecke, Polygone
  - Linien
    - gerade
    - gebogen
  - Linienzüge
- 3D
  - (Ober)-Flächen
    (2D Objekte in 3D)
    - eben
    - gekrümmt
  - Volumina
    - Kugel, Ellipsoid, ...
    - Zylinder, Torus, ...
    - Würfel, Quader, ...
  - Linien
    - meist Zylinder-förmig

# Graphische Elemente

## Eigenschaften

- Punkte (Pixel)
  - Farbe
    - Farbwert (hue)
    - Sättigung (saturation)
    - Helligkeit (brightness, lightness)
- Ebene geometrische Figuren
  - Farbe
  - Form
  - Orientierung
  - Größe
  - Textur
- Linien
  - Farbe
  - Form
    - Gerade
    - Linienzug (Polyline)
    - Gebogen
  - Breite
  - (Länge: meist durch Objekte, die verbunden werden, gegeben)
  - Stil
    - Durchgezogen
    - Strich-Punkt
    - Gepunktet

# Graphische Elemente

## Einschränkungen

- Farbe
  - Gleichzeitige Verwendung von Sättigung und Helligkeit nicht möglich
  - In 3D wird die Farbe durch das Beleuchtungsmodell bestimmt
    - Sättigung
    - Helligkeit
- Form, Orientierung und Textur (Stil von Linien)
  - bedingen Mindestgröße für die Erkennbarkeit
- Größe (Breite von Linien)
  - muss begrenzt sein (Minimum, Maximum)
  - gegebenenfalls muss der Wertebereich skaliert werden
- Textur sollte keinen visuellen Stress erzeugen

# Graphische Elemente

Eignung von graphischen Eigenschaften von graphischen Elementen für Datentypen

| Attribute | Nominal | Geordnet | Quantitativ |
|---|---|---|---|
| Farbwert | ✓ | × | × |
| Farbintensität | × | ✓ | ○ |
| Form (Stil) | ✓ | × | × |
| Orientierung | × | ○ | ○ |
| Größe (1D) | × | ○ | ✓ |
| Größe (2D) | × | ○ | ○ |
| Textur | ✓ | ○ | ○ |

# Graphische Elemente

Eignung von weiteren graphischen Eigenschaften für Datentypen

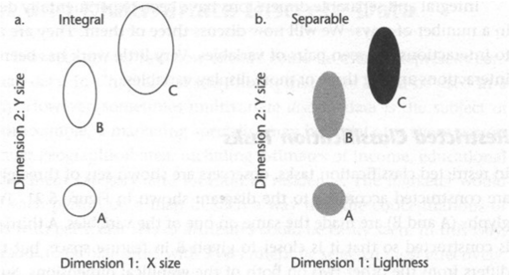
| Attribute | Nominal | Geordnet | Quantitativ |
|---|---|---|---|
| Position | ○ | ✓ | ✓ |

# Graphische Elemente: Glyphen

- Ein Glyph ist graphisches Objekt zur Repräsentation eines multivariaten Datenobjektes
- Kombinationen von graphischen Dimensionen (Variablen) lassen sich unterscheiden in
  - separable
  - integrale
- Ob graphische Variablen gemeinsam (integral) oder getrennt (separabel) wahrgenommen werden, muss durch Experimente ermittelt werden

# Graphische Elemente: Glyphen

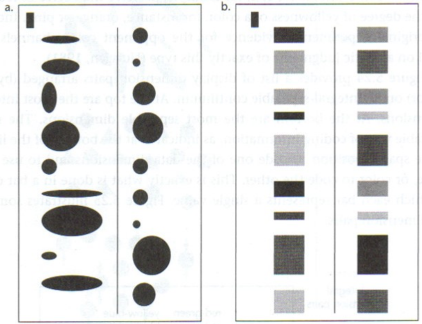
- Integrale Dimensionen
  - Breite und Höhe
  - Rot und Grün
    (ergeben zusammen Gelb)
  - ...
- Trennbare Dimensionen
  - Durchmesser und Farbe
  - ...




(a) The width and height of an ellipse are perceived integrally; therefore, B and C are perceived as more similar. (b) The gray value and the height of an ellipse are perceived as separable; therefore, A and B, which have identical lightness, are perceived as more similar.

# Graphische Elemente: Glyphen

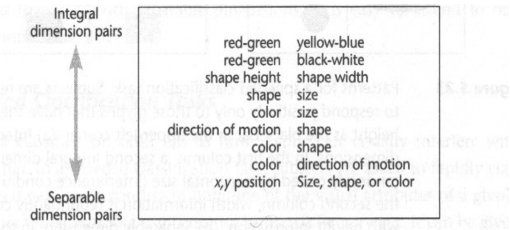
- Integrale Dimensionen
  - Breite und Höhe
  - Rot und Grün
    (ergeben zusammen Gelb)
  - ...
- Trennbare Dimensionen
  - Durchmesser und Farbe
  - ...




Patterns for a speeded classification task. Subjects are required to respond positively only to those glyphs that have the same height as the black bar in the upper-left corner. (a) Integral dimensions. In the first column, a second integral dimension is randomly coded by horizontal size (interference condition). In the second column, width information is redundantly coded with height information. (b) Separable dimension. In the first column, gray information is not correlated with height. In the second column, gray level is a redundant code.

# Graphische Elemente: Glyphen

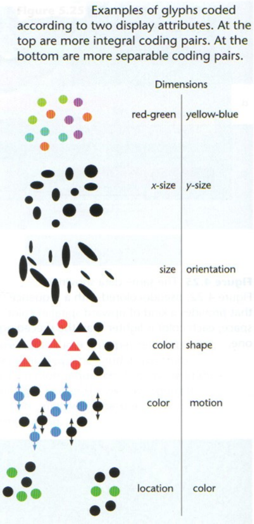
- Integrale Dimensionen
  - Breite und Höhe
  - Rot und Grün
    (ergeben zusammen Gelb)
  - ...
- Trennbare Dimensionen
  - Durchmesser und Farbe
  - ...
- Diese Unterscheidung ist nicht strikt;
  es gibt Übergänge




This table lists some of the display dimension pairs ranked in order from highly integral to highly separable.

# Graphische Elemente: Glyphen

- Integrale Dimensionen
  - Breite und Höhe
  - Rot und Grün
    (ergeben zusammen Gelb)
  - ...
- Trennbare Dimensionen
  - Durchmesser und Farbe
  - ...
- Diese Unterscheidung ist nicht strikt;
  es gibt Übergänge



# Graphische Elemente: Glyphen

- Multidimensionale Daten
  - 8 Dimensionen
  - Mindestens 2 Bit pro Dimension
- Nur trennbare Variablen können verwendet werden
  - 5 Dimensionen
  - 32 verschiedene Glyphen
- Diese können von visueller Vorverarbeitung schnell unterschieden werden

| Visual variable | Dimensionality | Comment |
|---|---|---|
| Spatial position of glyph | 3 dimensions: *X*, *Y*, *Z*. | |
| Color of glyph | 3 dimensions: defined by color opponent theory. | Luminance contrast is needed to specify all other graphical attributes. |
| Shape | 2–3? Dimensions unknown. | The dimensions of shape that can be rapidly processed are unknown. However, evidence suggests that size and degree of elongation are two primary ones. |
| Orientation | 3 dimensions: corresponding to orientation about each of the primary axes. | Orientation is not independent of shape. One object can have rotation symmetry with another. |
| Surface texture | 3 dimensions: orientation, size, and contrast. | Not independent of shape or orientation. Uses up one color dimension. |
| Motion coding | 2–3? Dimensions largely unknown, but phase may be useful. | |
| Blink coding: The glyph blinks on and off at some rate. | 1 dimension. | Motion and blink coding are highly interdependent. |

Graphical attributes that may be used in glyph design.

# Graphische Elemente

- Auswahl der visuellen Abbildung in Abhängigkeit von der Dimension der Daten
  - Dimension: Anzahl der Attribute (Variablen)

| Dimension | Bezeichung |
|---|---|
| 1D | Uni-Variate Daten |
| 2D | Bi-Variate Daten |
| 3D | Tri-Variate Daten |
| $\geq$4D | Multi-Variate Daten |
| | Hyper-Variate Daten |

- Darstellung der Daten in einem 2- oder 3-dimensionalen visuellen Raum
- Im Folgenden sind die Daten meistens quantitativ
  - Uni-, Bi- und Tri-Variate Daten: Position von Markierungen auf orthogonalen Achsen dargestellt
  - Multi-Variate Daten: Darstellung schwieriger
- Wahrnehmung ist sehr wichtig für
  - die visuelle Abbildung
  - die Bildung von visuellen Strukturen

# Uni-Variate Daten (1D)

Betrachtung eines Attributes

- Gegeben
  - $n$ Datenpunkte
  - 1 Attribut: $a_1$
  - Werte des Attributes $X = \{x_i | 1 \leq i \leq n\}$

| ID | Attribut $a_1$ |
|---|---|
| $id_1$ | $x_{1,1}$ |
| $id_2$ | $x_{2,1}$ |
| ... | |
| $id_n$ | $x_{n,1}$ |

- Ziel
  - Verständnis der Verteilung der Werte dieses Attributes
- Definitionen
  - $x_{min} = \min(X)$
  - $x_{max} = \max(X)$
- Beachte
  - Die $x_i$ müssen nicht geordnet sein

# Uni-Variate Daten (1D)

## Häufigkeitsverteilung

- Sei $D = \{d_j\}$ die Menge der **verschiedenen** Werte desselben Attributes
- Dann gilt
  - $D = \{d_j\} \subseteq \{x_i\} = X$
  - $n_d := |\{d_j\}| \leq |\{x_i\}| = n$
- Absolute Häufigkeit

$$h^a(d_j) := |\{x_i | x_i = d_j\}|$$

  Zähle für jeden Wert $d_j$, wie häufig er vorkommt
- Es gilt

$$\sum_{j=1}^{n_d} h^a(d_j) = n$$

# Uni-Variate Daten (1D)

## Häufigkeitsverteilung

- Sei $D = \{d_j\}$ die Menge der **verschiedenen** Werte desselben Attributes
- Absolute Häufigkeit

$$h^a(d_j) := |\{x_i | x_i = d_j\}|$$

  Zähle für jeden Wert $d_j$, wie häufig er vorkommt

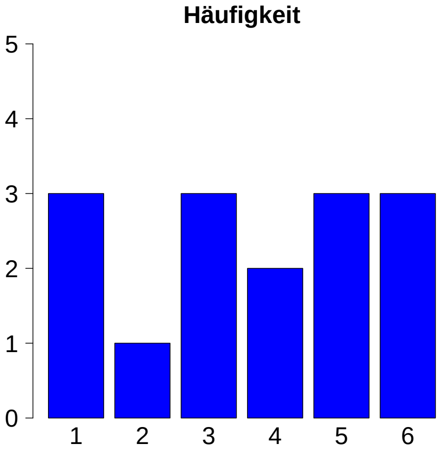
- Stelle die absolute Häufigkeit in einem Balkendiagramm dar
- Form: Rechteck
  - Breite: konstant
  - Höhe: $h^a(d_j)$
- Positionierung: Koordinatensystem
  - x-Achse: $j$
  - y-Achse: 0
- Farbe: konstant

# Uni-Variate Daten (1D)

## Häufigkeitsverteilung 1

| ID | Wert |
|---|---|
| 1 | 1 |
| 2 | 3 |
| 3 | 5 |
| 4 | 6 |
| 5 | 6 |
| 6 | 4 |
| 7 | 2 |
| 8 | 4 |
| 9 | 6 |
| 10 | 3 |
| 11 | 5 |
| 12 | 1 |
| 13 | 1 |
| 14 | 5 |
| 15 | 3 |

| Wert | Häufigkeit |
|---|---|
| 1 | 3 |
| 2 | 1 |
| 3 | 3 |
| 4 | 2 |
| 5 | 3 |
| 6 | 3 |

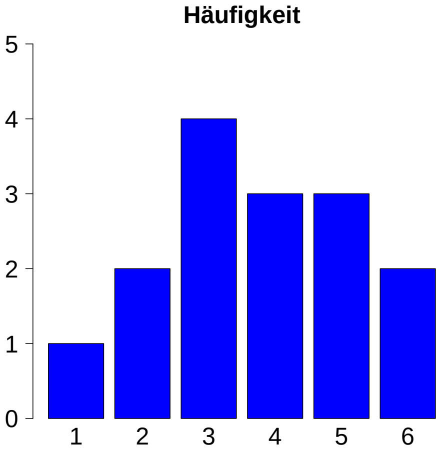


# Uni-Variate Daten (1D)

## Häufigkeitsverteilung 2

| ID | Wert |
|---|---|
| 1 | 1 |
| 2 | 3 |
| 3 | 5 |
| 4 | 6 |
| 5 | 2 |
| 6 | 4 |
| 7 | 3 |
| 8 | 4 |
| 9 | 6 |
| 10 | 3 |
| 11 | 5 |
| 12 | 4 |
| 13 | 3 |
| 14 | 5 |
| 15 | 3 |

| Wert | Häufigkeit |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 4 |
| 4 | 3 |
| 5 | 3 |
| 6 | 2 |



# Uni-Variate Daten (1D)

Balkendiagramm: Datentypen

- Kategorisch:
  - Reihenfolge unerheblich
- Geordnet:
  - Reihenfolge muss beachtet werden

$$\forall i < j : d_i < d_j$$

- Numerisch Diskret:
  - Fehlende Werte müssen mit Häufigkeit 0 angezeigt werden

$$\forall d_i : (d_{min} < d_i < d_{max} \land d_i \notin D) \longrightarrow h^a(d_i) = 0$$

- Numerisch Kontinuierlich:
  - Abstand zwischen zwei Werten muss gleich sein
  - Besser: Verwendung von Histogrammen

# Uni-Variate Daten (1D)

Balkendiagramm: Höhe der Balken

- Maximale Häufigkeiten größer als die Anzahl der zur Verfügung stehenden Pixel der Zeichnung
  - Höhe des Balkens für kleine Häufigkeiten kleiner als 1
- Um immer zwischen Häufigkeit 0 und kleinen Häufigkeiten unterscheiden zu können
  - Die minimale Höhe eines Balkens für Häufigkeiten grösser 0 auf mindestens 1 Pixel setzen

# Uni-Variate Daten (1D)

Häufigkeitsverteilung

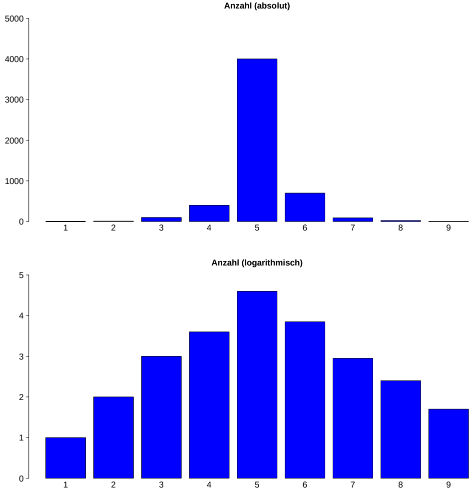
- Logarithmische Skalen
  - Umfassen die Häufigkeiten mehrere Größenordnungen, so kann eine logarithmische Transformation hilfreich sein

$$h(d_i) = \begin{cases} \log_b(h^a(d_i)) + 1 & h^a(d_i) > 0 \\ 0 & h^a(d_i) = 0 \end{cases}$$

  - Meist: $b = 10$

# Uni-Variate Daten (1D)

| Wert | Häufigkeit (absolut) | Häufigkeit (logarithmisch) |
|---:|---:|---:|
| 1 | 1 | 1,00 |
| 2 | 10 | 2,00 |
| 3 | 100 | 3,00 |
| 4 | 400 | 3,60 |
| 5 | 4000 | 4,60 |
| 6 | 700 | 3,85 |
| 7 | 90 | 2,95 |
| 8 | 25 | 2,40 |
| 9 | 5 | 1,70 |



# Uni-Variate Daten (1D)

## Häufigkeitsverteilung

- Relative Häufigkeit

$$h^r(d_j) := \frac{h^a(d_j)}{n}$$

- Eigenschaften
  - $0 \leq h^r(d_j) \leq 1$
  - $\sum_{j=1}^{n_d} h^r(d_j) = 1$

- Prozentuale Häufigkeit

$$h^p(d_j) := h^r(d_j) \cdot 100$$

- Eigenschaften
  - $0 \leq h^p(d_j) \leq 100$
  - $\sum_{j=1}^{n_d} h^p(d_j) = 100$

# Uni-Variate Daten (1D)

Häufigkeitsverteilung

- Basis (Bezugspunkt) der Relativierung muss deutlich sein
- Was ist wichtiger: absoluter oder relativer Wert
- Bei sehr kleinem $n$ sind relative Werte in der Regel irreführend
  - $n = 5$
  - Differenz absolut: 1
    $\rightarrow$ Differenz prozentual: 20

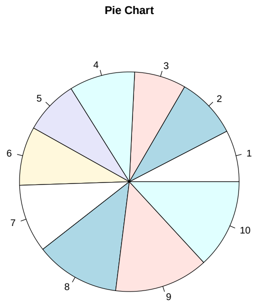
| $h^a(d_j)$ | $h^p(d_j)$ |
|---:|---:|
| 0 | 0% |
| 1 | 20% |
| 2 | 40% |
| 3 | 60% |
| 4 | 80% |
| 5 | 100% |

# Uni-Variate Daten (1D)

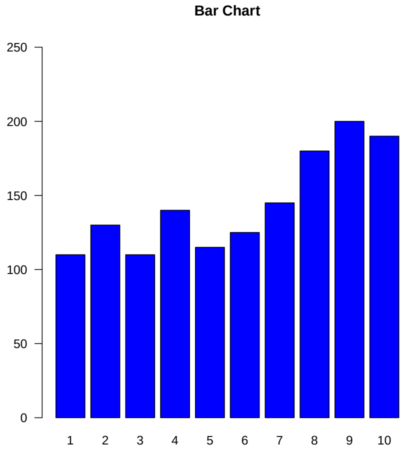
Pie Chart

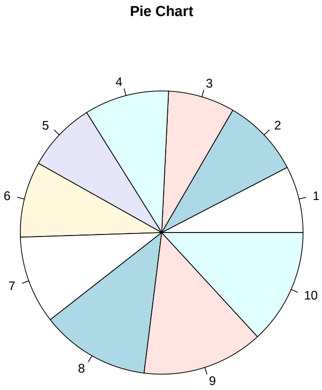
- Zeigen relative/anteilige Verteilung (meist: prozentuale Verteilung)
- Meist für Geschäftsgraphiken
- Kaum Verwendung im wissenschaftlichen Bereich
- Probleme
  - Fläche und Winkel sind schwieriger zu interpretieren als Länge
  - Schwierig für (numerische) Vergleiche
  - Nutzung von vielen Pie-Charts gleichzeitig ist sehr schwierig

| | |
|---|---|
| 1 | 110 |
| 2 | 130 |
| 3 | 110 |
| 4 | 140 |
| 5 | 115 |
| 6 | 125 |
| 7 | 145 |
| 8 | 180 |
| 9 | 200 |
| 10 | 190 |



# Uni-Variate Daten (1D)



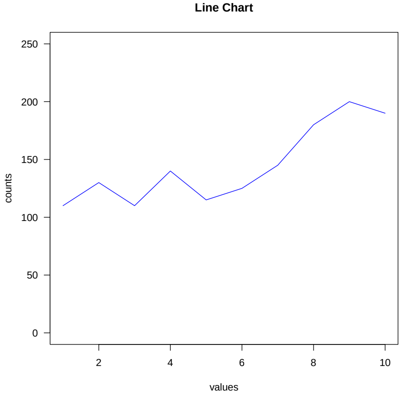
# Uni-Variate Daten (1D)

Line Chart

- Zeigt Form und Verlauf der Verteilung
- Zeigt nicht die Werte
- Geeignet für
  - Kontinuierliche Daten
  - Zeitreihen
- Nicht geeignet für
  - Kategorische Daten
  - Geordnete Daten
  - Diskrete Daten

  (erweckt Eindruck, dass Zwischenwerte existieren)

| | |
|---|---|
| 1 | 110 |
| 2 | 130 |
| 3 | 110 |
| 4 | 140 |
| 5 | 115 |
| 6 | 125 |
| 7 | 145 |
| 8 | 180 |
| 9 | 200 |
| 10 | 190 |

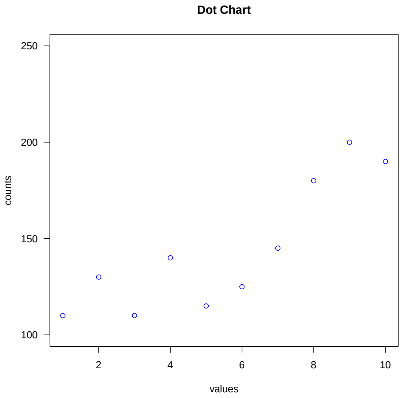


# Uni-Variate Daten (1D)

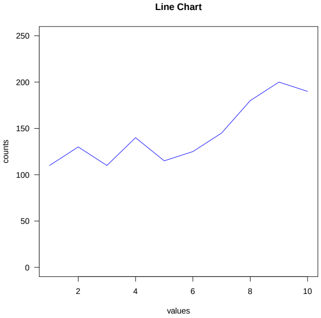
## Dot Chart

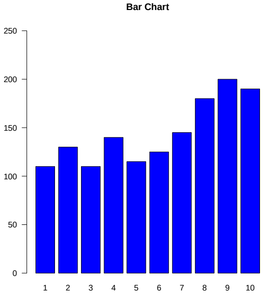
- Punkte anstelle von Balken
- Geeignet für
  - Kategorische Daten
  - Geordnete Daten
  - Diskrete Daten
  - Kontinuierliche Daten
  - Zeitreihen
- Sinnvoll, falls y-Achse nicht bei 0 beginnt

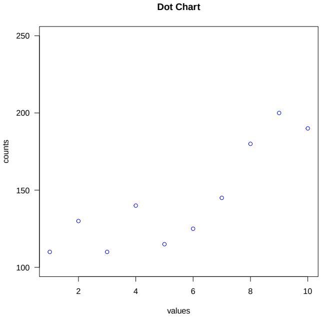
| | |
|---|---|
| 1 | 110 |
| 2 | 130 |
| 3 | 110 |
| 4 | 140 |
| 5 | 115 |
| 6 | 125 |
| 7 | 145 |
| 8 | 180 |
| 9 | 200 |
| 10 | 190 |



# Uni-Variate Daten (1D)



# Uni-Variate Daten (1D)

## Häufigkeitsverteilung: Klasseneinteilung

- Gruppierung der $d_i \in D$
  - $n_k$: Anzahl der Klassen
  - $w = \frac{d_{max} - d_{min}}{n_k}$: Breite der Klassen
  - $$\forall\, 0 \leq j \leq n_k - 1 : k_j = \{d_i \in D | d_{min} + w \cdot j \leq d_i < d_{min} + w \cdot (j+1)\}$$
  - $$h^*(k_j) = \sum_{d_i \in k_j} h^*(d_i)$$
- Insbesondere bei kontinuierlichen Werten sinnvoll

# Uni-Variate Daten (1D)

Häufigkeitsverteilung: Klasseneinteilung

- Kategorische Werte
  - Zusammenfassung abhängig von den Kategorien
- Numerische Werte
  - Gleich große Intervalle
  - Anzahl der Klassen

$$n_k = \begin{cases} \sqrt{n_d} & n_d \leq 1000 \\ 10 \cdot log(n_d) & n_d > 1000 \end{cases}$$

  - Anzahl der Unterteilungen soll Verlauf widerspiegeln
- Kontinuierliche Werte
  - Intervalle rechts oder links offen

# Uni-Variate Daten (1D)

## Häufigkeitsverteilung: Klasseneinteilung

- Darstellung mit Histogrammen
  - Balkendiagramm mit Klassen und deren Häufigkeiten
- Form: Rechteck
  - Breite: konstant
  - Höhe: $h^*(k_j)$
- Positionierung: Koordinatensystem
  - x-Achse: $j$
  - y-Achse: 0
- Farbe: konstant

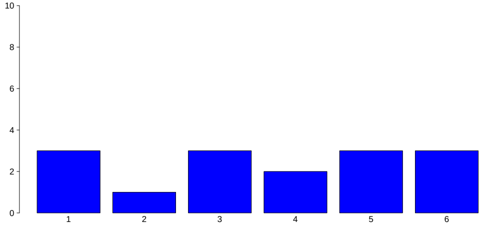
- Bemerkungen
  - Falls
    - $\forall d_j \in D : d_j > 0$
    - und die Werte umfassen mehrere Größenordnungen

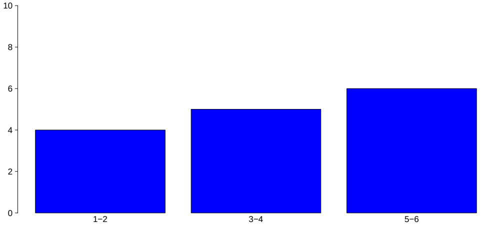
    dann kann es sinnvoll sein,
    - den Logarithmus der $d_j$ zu berechnen
    - und die Klasseneinteilung auf den logarithmischen Werten vorzunehmen
  - Falls $\exists j_0 : d_{j_0} = 0$:
    füge eine Klasse mit dem Wert 0 und Häufigkeit $h^*(d_{j_0})$ hinzu

# Uni-Variate Daten (1D)

| Wert | Häufigkeit |
|---|---|
| 1 | 3 |
| 2 | 1 |
| 3 | 3 |
| 4 | 2 |
| 5 | 3 |
| 6 | 3 |

| Wert | Häufigkeit |
|---|---|
| 1 – 2 | 4 |
| 3 – 4 | 5 |
| 5 – 6 | 6 |

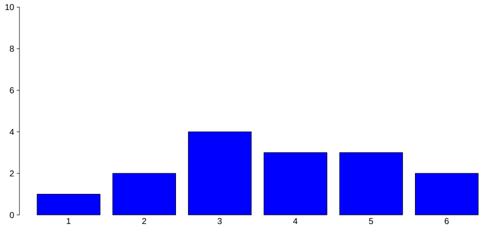


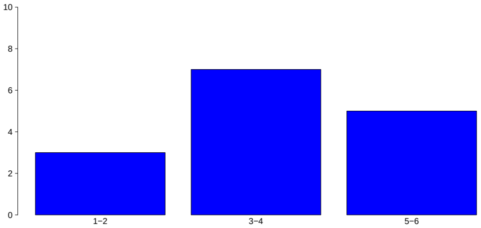



# Uni-Variate Daten (1D)

| Wert | Häufigkeit |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 4 |
| 4 | 3 |
| 5 | 3 |
| 6 | 2 |

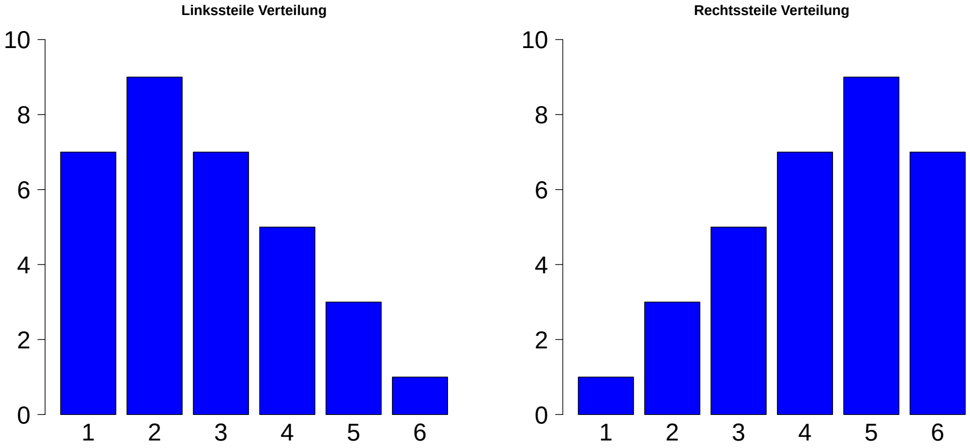
| Wert | Häufigkeit |
|---|---|
| 1 – 2 | 3 |
| 3 – 4 | 7 |
| 5 – 6 | 5 |



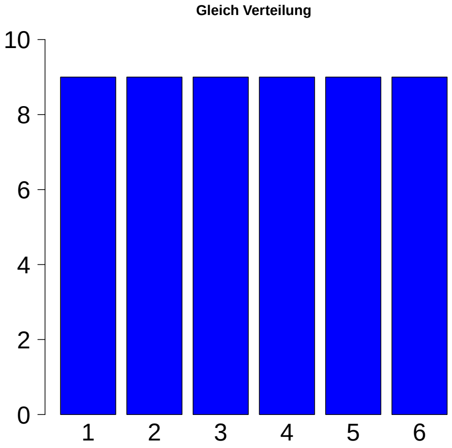
# Uni-Variate Daten (1D)

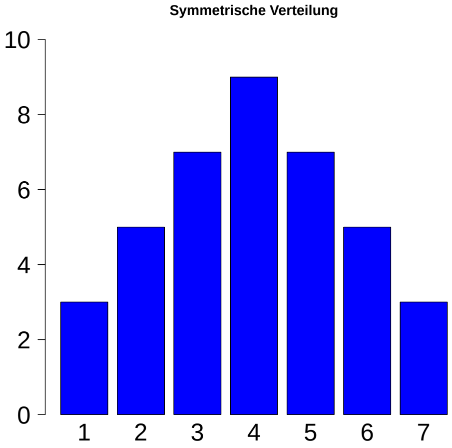
## Verteilungen



# Uni-Variate Daten (1D)

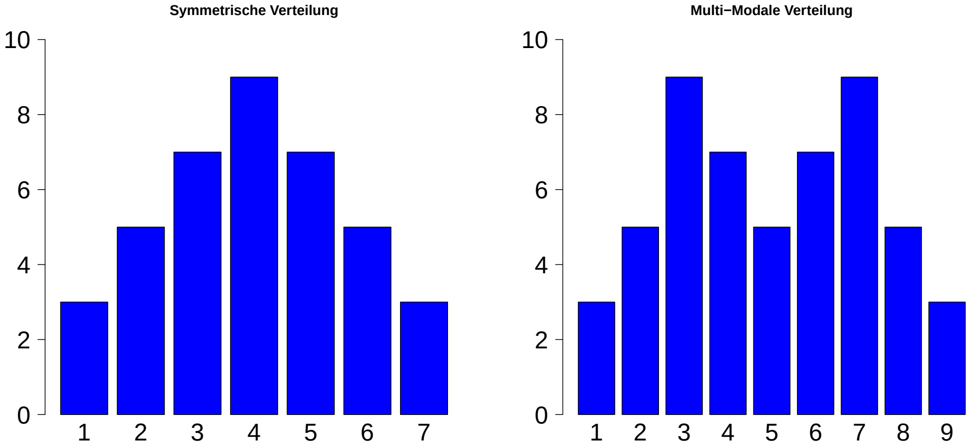
## Verteilungen



# Uni-Variate Daten (1D)

## Verteilungen



# Uni-Variate Daten (1D): Beschreibende Statistik

- Ziel: Reduktion
- Viele Zahlen → wenige Zahlen

| Zentrale Tendenz | Variabilität |
|---|---|
| Modus | |
| Median | Durchschnittliche absolute Abweichung |
| Mittelwert | Varianz, Standardabweichung |

- Kategorische Werte
  - Modus
- Geordnete Werte zusätzlich
  - Median
  - Durchschnittliche absolute Abweichung
- Kontinuierliche Werte zusätzlich
  - Mittelwert
  - Varianz
  - Standardabweichung

# Uni-Variate Daten (1D): Beschreibende Statistik

- Modus (mode)
  - Häufigster Wert:

$$modus = \{d_j | h^*(d_j) = \max_{1 \leq i \leq d_n} \{h^*(d_i)\}\}$$

  - Falls eindeutig:
    uni-modale Verteilung
  - Falls nicht eindeutig:
    multi-modale Verteilung

# Uni-Variate Daten (1D): Beschreibende Statistik

- Median (median)
  - Mittlerer Wert
  - $$median = \begin{cases} x_i & i = \frac{n+1}{2} \quad n \text{ ungerade} \\ \frac{x_i + x_{i+1}}{2} & i = \frac{n}{2} \quad n \text{ gerade} \end{cases}$$
  - ist gleich dem 2.ten Quartil

- Durchschnittliche absolute Abweichung
  - $$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - median|$$
  - Es gilt:
    $$\sum_{i=1}^{n} |x_i - m| \text{ ist minimal}$$
    $$\Updownarrow$$
    $$m = median$$

# Uni-Variate Daten (1D): Beschreibende Statistik

- Population
  - Menge aller möglichen Elemente, die gemessen werden sollen $\{X_i\}$
  - Theoretische Konstruktion
  - Größe der Population: $N = |\{X_i\}|$
- Messwerte
  - Menge aller Elemente, die tatsächlich gemessen wurden $\{x_i\}$
  - Teilmenge der zugehörigen Population
  - Anzahl der Messwerte: $n = |\{x_i\}|$

| | Population | Messwerte |
|---|---|---|
| Mittelwert | $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$ | $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ |
| Varianz | $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$ | $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$ |
| Standardabweichung | $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$ | $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$ |

# Uni-Variate Daten (1D): Beschreibende Statistik

## Mittelwert (mean)

- Alternative Berechnung

$$\begin{aligned}
\overline{x} &= \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \\
&= \frac{1}{n} \cdot \sum_{j=1}^{n_d} \cdot h^a(d_j) \cdot d_j \\
&= \sum_{j=1}^{n_d} \frac{1}{n} \cdot h^a(d_j) \cdot d_j \\
&= \sum_{j=1}^{n_d} h^r(d_j) \cdot d_j
\end{aligned}$$

- Es gilt
  - $$\sum_{i=1}^{n} (x_i - \overline{x}) = 0$$
  - $$\sum_{i=1}^{n} (x_i - \overline{x})^2 \text{ ist minimal}$$
    $$\Updownarrow$$
    $$\overline{x} = mean$$

# Uni-Variate Daten (1D): Beschreibende Statistik

- Anzahl der Freiheitsgrade
  - Gegeben:
    - $n$ Beobachtungen
    - $m$ Mittelwerte
  - Frage: Wie viele Beobachtungen können variieren?
  - Antwort: $n - m$
  - Begründung für $m = 1$:
    Wegen $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$
    können nur $n - 1$ der $n$ Variablen $x_i$
    frei gewählt werden
  - Daraus ergibt sich der Divisor für $s^2$

# Uni-Variate Daten (1D): Beschreibende Statistik

Bemerkungen

- Mittelwert, Varianz, Standardabweichung
  - Empfindlich gegen Ausreißer
  - Keine (geringe) Aussagekraft bei nicht-symmetrischen Verteilungen
- Median
  - Begrenzte Aussagekraft bei multi-modalen Verteilungen

| Verteilung | |
|---|---|
| Links-seitig | $Modus \leq Median \leq Mittelwert$ |
| Symmetrisch | $Modus = Median = Mittelwert$ |
| Rechts-seitig | $Modus \geq Median \geq Mittelwert$ |

# Uni-Variate Daten (1D): Beschreibende Statistik

- Symmetrie
  - Schiefe (skewness)

$$schiefe = \frac{1}{n} \cdot \sum_{i=1}^{n} \left( \frac{x_i - \overline{x}}{s} \right)^3$$

| Schiefe | Verteilung |
|---|---|
| $\approx 0$ | Symmetrisch |
| $\gg 0$ | Links-seitig |
| $\ll 0$ | Rechts-seitig |

- Vergleich mit der Normalverteilung
  - Wölbung (kurtosis)

$$wölbung = \frac{1}{n} \cdot \sum_{i=1}^{n} \left( \frac{x_i - \overline{x}}{s} \right)^4$$

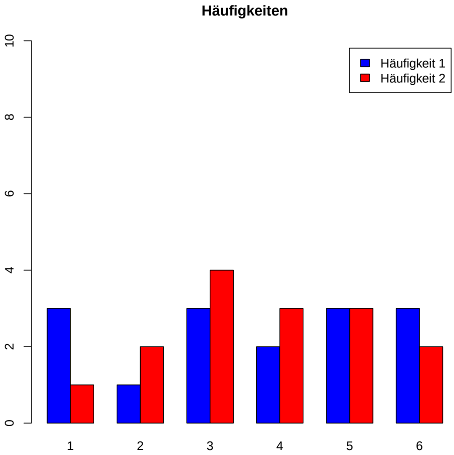
  - Exzess

$$exzess = wölbung - 3$$

| Exzess | Verteilung |
|---|---|
| $\approx 0$ | Normalgipflig |
| $\gg 0$ | Steilgipflig (schmaler, höher) |
| $\ll 0$ | Flachgipflig (breiter, flacher) |

# Uni-Variate Daten (1D)

## Vergleich mehrerer Verteilungen

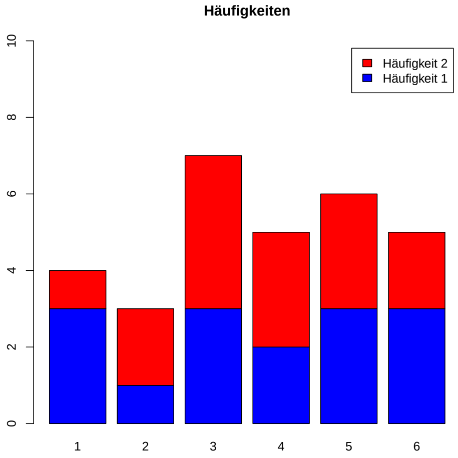



- Bemerkungen
  - gut:
    Vergleich zwischen Häufigkeit 1 und Häufigkeit 2 pro $d_j$
  - nicht immer optimal:
    Vergleich der $d_j$ innerhalb Häufigkeit 1 oder Häufigkeit 2
    - Insbesondere, falls kleine Unterschiede bei großem Wertebereich der Häufigkeiten auftreten

# Uni-Variate Daten (1D)

## Vergleich mehrerer Verteilungen

- Alternative: **Stacked Bar-Chart**
  - Gegeben: mehrere Attribute mit Häufigkeiten
  - Nur sinnvoll, falls Aggregation der Häufigkeiten im Vordergrund steht
    - Häufigkeit 1: Vergleich einfach
    - Häufigkeit 2: Vergleich nicht offensichtlich
  - Ansonsten: betrachte aggregierten Wert als eigenes Attribut



# Uni-Variate Daten (1D)

Normalverteilung

- Dichtefunktion

$$f(x,\mu,\sigma) = \frac{1}{\sqrt{2\cdot\pi\cdot\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\cdot\sigma^2}}$$

  - Mittelwert, Median, Modus: $\mu$
  - Varianz: $\sigma^2$
  - Schiefe: 0
  - Wölbung: 3
  - Exzess: 0
- Standardnormalverteilung
  - $\mu = 0,\ \sigma = 1$

$$f(x) = \frac{1}{\sqrt{2\cdot\pi}} \cdot e^{-\frac{x^2}{2}}$$

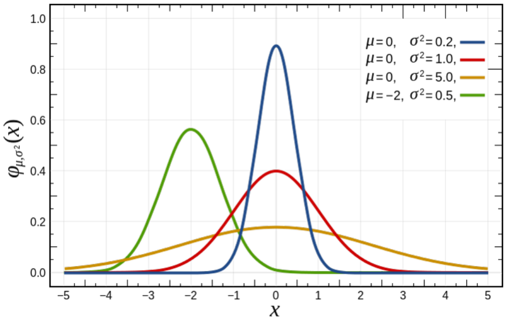



Abbildung: [Wikipedia]

# Uni-Variate Daten (1D)

## Poisson-Verteilung

- Dichtefunktion
  $$f(k,\lambda) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$
  - Mittelwert: $\lambda$
  - Median: $\lambda - \ln(2) \leq median \leq \lambda + \frac{1}{3}$
  - Varianz: $\lambda$
  - Schiefe: $\frac{1}{\sqrt{\lambda}}$
  - Wölbung: $3 + \frac{1}{\lambda}$
- Für kleine $\lambda$ stark asymmetrisch
- Für $\lambda \geq 30$
  - Annähernd normalverteilt
    $$(\mu = \lambda,\ \sigma^2 = \lambda)$$

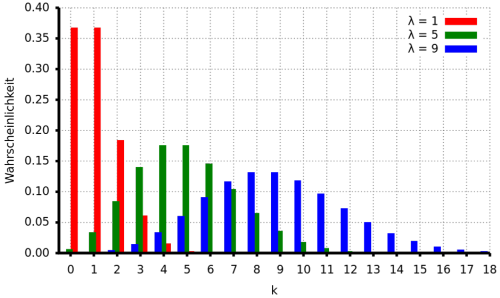



Abbildung: [Wikipedia]

# Uni-Variate Daten (1D)

Exponentialverteilung

- Dichtefunktion
$$f(x,\lambda) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$
  - Mittelwert: $\frac{1}{\lambda}$
  - Median: $\frac{\ln(2)}{\lambda}$
  - Varianz: $\frac{1}{\lambda^2}$
  - Schiefe: 2
  - Wölbung: 9

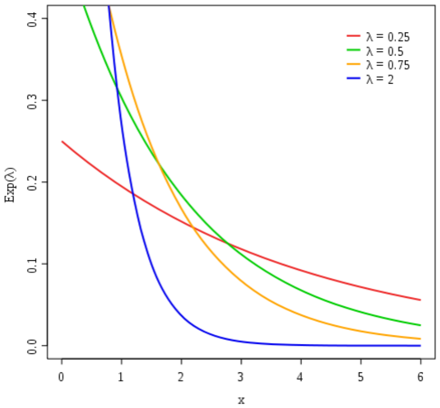



Abbildung: [Wikipedia]

# Uni-Variate Daten (1D)

Kummulierte Häufigkeiten

- Voraussetzung: geordnete Daten
- Frage: Wie viele Werte sind kleiner als der aktuelle Wert?
- Seien die Werte $d_j$ geordnet: $\forall i < j : d_i \leq d_j$
- Summiere die Häufigkeiten vom kleinsten bis zum aktuellen Datenpunkt auf

$$H^*(d_k) = \sum_{j=1}^{k} h^*(d_j)$$

- $$d_{p\%} := \max_{1 \leq j \leq n_d} \{d_j | H^*(d_j) \leq p\%\}$$

# Uni-Variate Daten (1D)

## Quantile

- Gliederung der Verteilung in Teilbereiche
- Das Quantil ist der größte Wert $d_i$ unterhalb dessen höchstens $p\%$ der Werte liegen
- Typisch: Quartile $Q_i$
  - 4 Teilbereiche ($n = 4$)
  - Quartil $Q_1$: bis zu 25% der Werte
  - Quartil $Q_2$: bis zu 50% der Werte
  - Quartil $Q_3$: bis zu 75% der Werte
  - Quartil $Q_4$: 100% der Werte
- Bemerkung
  - Quartil $Q_2$ entspricht dem Median

# Uni-Variate Daten (1D)

Boxplot

- Median:
  - Darstellung: Linie
- $[d_{25\%}, d_{75\%}]$
  - Interquartile-Range: $IQR = d_{75\%} - d_{25\%}$
  - Darstellung: Box

- Tukey-Boxplot
  - Whisker
    - Oberer Whisker: $d_{75\%} + 1,5 \cdot IQR$
    - Unterer Whisker: $d_{25\%} - 1,5 \cdot IQR$
    - Darstellung: Balken
  - Ausreißer
    - Darstellung: individuelle Datenpunkte
    - Bis $3 \cdot IQR$: “milde Ausreißer”

- Min-Max-Boxplot
  - Whisker
    - Oberer Whisker: $d_{max}$
    - Unterer Whisker: $d_{min}$
    - Darstellung: Balken
  - Ausreißer
    - Darstellung: keine

# Uni-Variate Daten (1D)

| Werte |
|---|
| 15 |
| 20 |
| 25 |
| 30 |
| 33 |
| 34 |
| 35 |
| 37 |
| 40 |
| 75 |

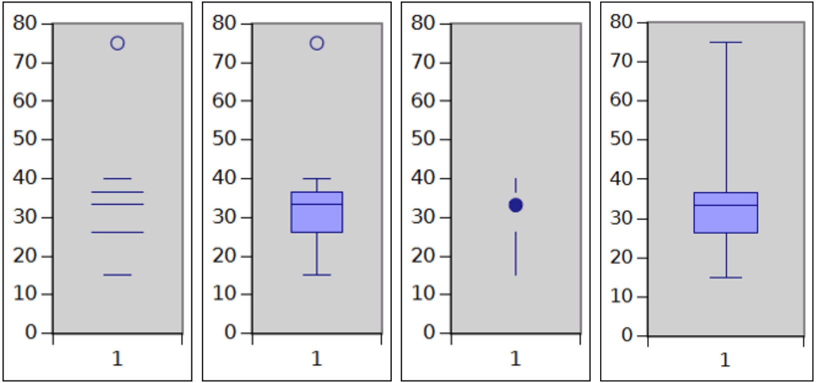



Abbildung: Links: Tukey-Boxplot ohne senkrechte Linien; Mitte-Links: Tukey-Boxplot; Rechts: Alternative Darstellung des Tukey-Boxplots; Mitte-Rechts: Min-Max-Boxplot

# Bi-Variate Daten (2D)

- Gegeben
  - $n$ Datenpunkte
  - 2 Attribute: $a_1, a_2$
  - Werte der Attribute
    $X = \{x_{ij} | 1 \leq i \leq n \wedge 1 \leq j \leq 2\}$

| ID | Attribut $a_1$ | Attribut $a_2$ |
|---|---|---|
| $id_1$ | $x_{1,1}$ | $x_{1,2}$ |
| $id_2$ | $x_{2,1}$ | $x_{2,2}$ |
| $\dots$ | | |
| $id_n$ | $x_{n,1}$ | $x_{n,2}$ |

- Typische Fragen:
  - Welchen Zusammenhang gibt es zwischen den Attributen?
    - Korrelation (positive Korrelation)
    - Anti-Korrelation (negative Korrelation)
    - Keinen Zusammenhang
- Positive Korrelation

$$\forall 1 \leq i, j \leq n : x_{i,1} < x_{j,1} \rightarrow x_{i,2} \leq x_{j,2}$$

- Negative Korrelation

$$\forall 1 \leq i, j \leq n : x_{i,1} < x_{j,1} \rightarrow x_{i,2} \geq x_{j,2}$$

# Bi-Variate Daten (2D)

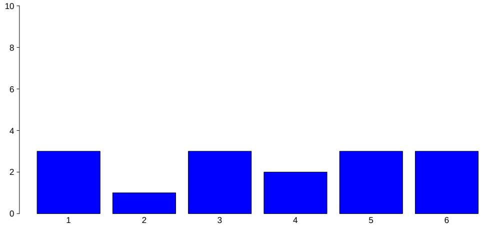
- Prinzipiell lassen sich mehrere Attribute mit den bisherigen Charts darstellen
- Vorgehen:
  - Wähle ein primäres Attribut
  - Alle anderen Attribute (“Häufigkeiten”) werden in Bezug auf das primäre Attribut dargestellt
- Vorteile
  - Gut für Vergleiche von Attributen gleicher Charakteristik
- Einschränkungen
  - Bezug immer nur zu einem Attribut
  - Häufigkeiten sollten in der gleichen Größenordnung liegen

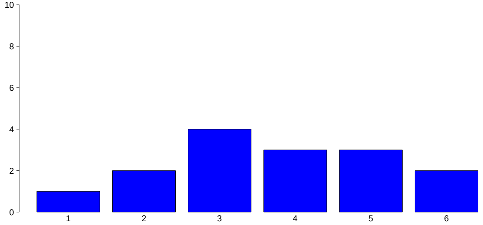
# Bi-Variate Daten (2D)

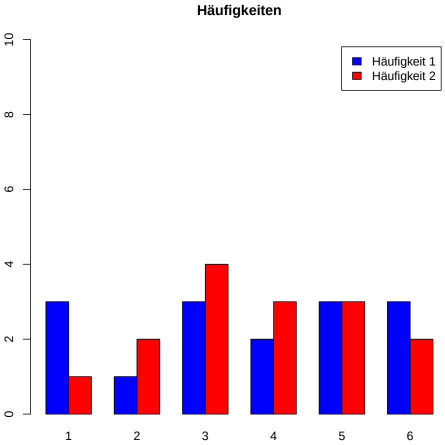
- Small-Multiples
  - Ein Diagramm pro Attribut
  - Mehrere Diagramme auf der gleichen Zeichenfläche
- Beispiel Pie-Chart
  - Mehrere Pie-Charts notwendig
  - Vergleich bereits innerhalb eines Pie-Charts schwierig

# Bi-Variate Daten (2D)

## Bar-Charts



# Bi-Variate Daten (2D)

## Gekoppelte Histogramme

- Englisch: linked histograms
- Kopplung durch
  - Farbe
  - Pattern (Textur)
  - Interaktion

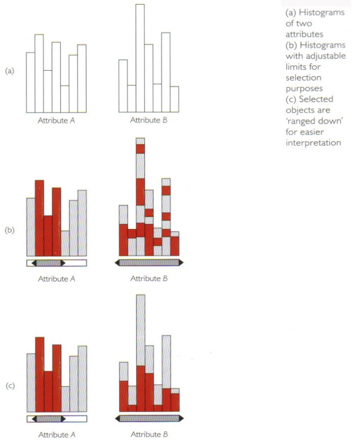



Abbildung: [Spence2001]

# Bi-Variate Daten (2D)

- Tatsächlich wurden schon bisher immer mindestens 2 Attribute betrachtet

| Attribut 1 | Attribut 2 |
|---|---|
| Wert | Häufigkeit pro Wert |
| Kategorie | Häufigkeit pro Kategorie |

- Wichtiges Unterscheidungsmerkmal für Attribut 1: Farbe
- Daraus folgt: Anzahl der gleichzeitig zu betrachtenden Werte oder Kategorien ist auf $\approx 6 - 10$ begrenzt

# Bi-Variate Daten (2D)

Scatterplot

- Gegeben:

$$X = \{x_{ij} | 1 \leq i \leq n \wedge 1 \leq j \leq 2\}$$

- Für zwei Attribute
  - Mindestens geordnet
  - Meist: numerisch
- Erlaubt es, eine Korrelation zwischen den beiden Attributen zu erkennen

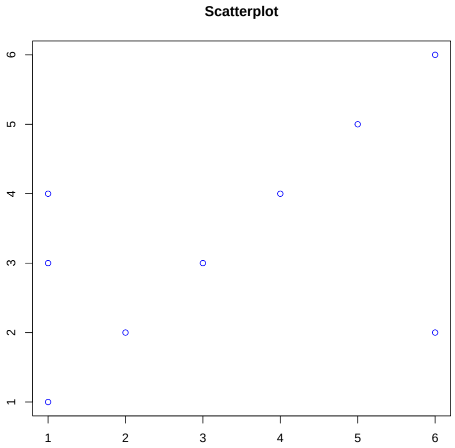
- Darstellung:
  - Position $(x, y) = (x_{i,1}, x_{i,2})$
  - Kreis an der Position $(x, y)$
- Zusätzlich
  - Koordinatenachsen
  - Eventuell Regressionsgerade

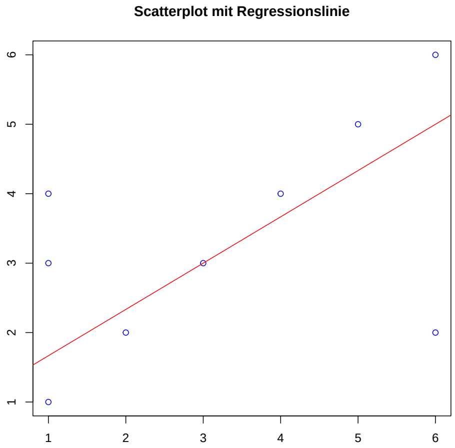
# Bi-Variate Daten (2D)

## Scatterplot

- Beachte:
  - Overplotting bei gleichen Werten
    - $(x_{i,1}, x_{i,2}) = (x_{j,1}, x_{j,2}), i \neq j$
  - Skalierung der Achsen
    - Äquidistant $\leftrightarrow$ Logarithmisch
  - Achsenkreuzung bei $(0, 0)$ oder nicht

# Bi-Variate Daten (2D)



# Tri-Variate Daten (2D)

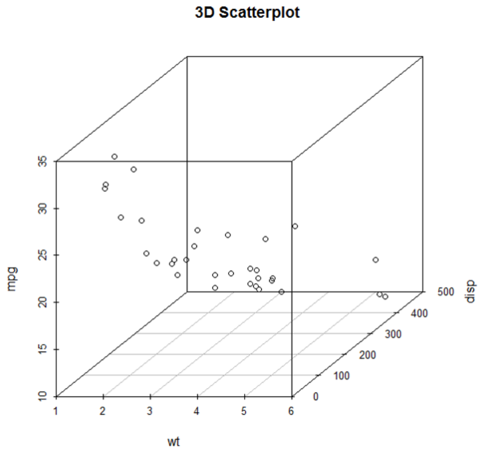
- Gegeben
  - $n$ Datenpunkte
  - 3 Attribute
  - Werte der Attribute $X = \{x_{ij} | 1 \leq i \leq n \wedge 1 \leq j \leq 3\}$
- Typische Fragen:
  - Welchen Zusammenhang gibt es zwischen den Attributen?
    - Korrelation
    - Anti-Korrelation
    - Keinen Zusammenhang

| ID | Attribut $a_1$ | Attribut $a_2$ | Attribut $a_3$ |
|---|---|---|---|
| $id_1$ | $x_{1,1}$ | $x_{1,2}$ | $x_{1,3}$ |
| $id_2$ | $x_{2,1}$ | $x_{2,2}$ | $x_{2,3}$ |
| ... | | | |
| $id_n$ | $x_{n,1}$ | $x_{n,2}$ | $x_{n,3}$ |

# Tri-Variate Daten (2D)

## 3D-Scatterplot

- Logische oder geometrische 2D-Projektion der 3D Darstellung
- Probleme:
  - 2D-Darstellung eines 3D-Raumes
  - Wie bestimmt man die Werte der Attribute aus der Darstellung?
  - Verdeckungsproblem
  - Datensatz: “Cars” (Autos)



# Tri-Variate Daten (2D)

Small Multiples: drei 2D-Scatterplots

- Betrachtung aller achsenparallelen Projektionen (2D-Scatterplots)
- Alle drei möglichen 1-1 Beziehungen können untersucht werden

**FIGURE 3.13** Projection of the points in Figure 3.12 onto all pairs of axes


Abbildung: [Spence2001]

# Tri-Variate Daten (2D)

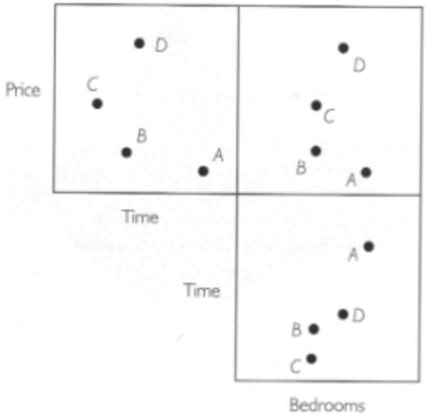



Abbildung: Scatterplot-Matrix: Übersichtliche Anordnung von Scatterplots [Spence2001]

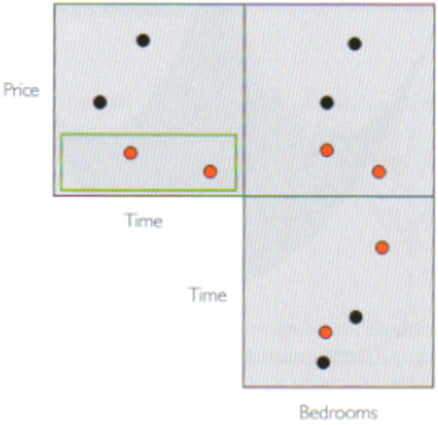



Abbildung: Brushing: Markierung gleicher Punkte [Spence2001]

# Tri-Variate Daten (2D)

Alternative Darstellung

- Dimensionen 1 & 2: Position
- Dimension 3: Größe oder Farbe
- Nachteil
  - Unterschiedliche Qualität in der Darstellung der Dimensionen macht es schwierig(er), Korrelationen zu sehen

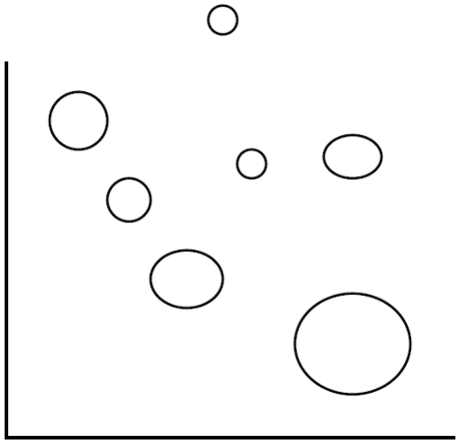


Abbildung: [Spence2001]

# Multi-Variate Daten

- Gegeben
  - $n$ Datensätze
  - $m > 3$ Attribute
  - Werte der Attribute $X = \{x_{ij} | 1 \leq i \leq n \wedge 1 \leq j \leq m\}$
- Typische Fragen:
  - Welchen Zusammenhang gibt es zwischen den Attributen?
    - Korrelation
    - Anti-Korrelation
    - Keinen Zusammenhang

| ID | Attribut $a_1$ | Attribut $a_2$ | Attribut $a_3$ | ... | Attribut $a_m$ |
|---|---|---|---|---|---|
| $id_1$ | $x_{1,1}$ | $x_{1,2}$ | $x_{1,3}$ | ... | $x_{1,n}$ |
| $id_2$ | $x_{2,1}$ | $x_{2,2}$ | $x_{2,3}$ | ... | $x_{2,n}$ |
| ... | | | | | |
| $id_n$ | $x_{n,1}$ | $x_{n,2}$ | $x_{n,3}$ | ... | $x_{n,m}$ |

# Multi-Variate Daten

- Problem:
  - Mehr als drei Attribute können nicht mehr nur über die Position repräsentiert werden

- Lösungen:
  - Übertragung der Tri-Variaten Ansätze
    - Scatterplot-Matrizen
    - Repräsentation mittels anderer visueller Attribute (z.B. Größe)
  - Alternative Methoden

# Multi-Variate Daten

- "Cars"-Datensatz

| | |
|---|---|
| Baujahr | Year |
| Beschleunigung | Acceleration |
| Leistung | Horsepower |
| Zylinder | Cylinders |
| Gewicht | Weight |
| Herstellungsland | Origin |
| Reichweite | MPG |

- Beispiele: [PWR2004]

# Multi-Variate Daten

1. Geometrische Ansätze (Projektion)
2. Achsenbasierte Ansätze
3. Heatmap

# Multi-Variate Daten

## Geometrische Ansätze (Projektion)

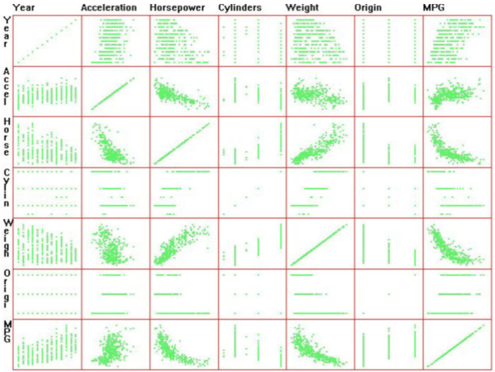
- Scatterplot-Matrix
  - Entspricht orthogonaler Projektion
- Hyperslice
  - Ohne strikte Festlegung auf orthogonale Projektionen
  - $n^2$ Schnitte fester Breite durch Daten legen
- Prosection Views
  - Auswahl einer $n$-dimensionalen Teilmenge (Hyperwürfel)
  - Wird bei Projektion mit anderer Farbe dargestellt

# Multi-Variate Daten: Scatterplot-Matrix

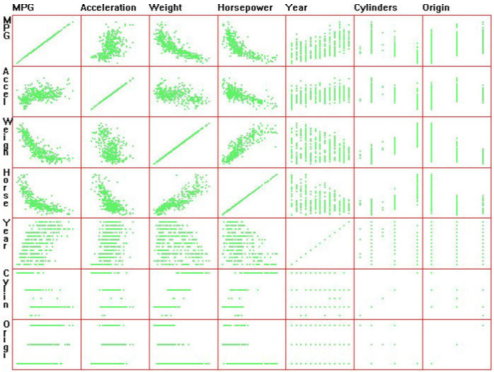
- Scatterplot:
  Korrelation zwischen 2 Attributen
- $m$ Attribute $\rightarrow$ $m \cdot m$ Attributpaare
- Bilde $m \times m$ Matrix von Scatterplots

| | $j_1$ | $j_2$ | ... | $j_m$ |
|---|---|---|---|---|
| $j_1$ | $j_1 \times j_1$ | $j_1 \times j_2$ | ... | $j_1 \times j_m$ |
| $j_2$ | $j_2 \times j_1$ | $j_2 \times j_2$ | ... | $j_2 \times j_m$ |
| ... | | | | |
| $j_m$ | $j_m \times j_1$ | $j_m \times j_2$ | ... | $j_m \times j_m$ |

# Multi-Variate Daten: Scatterplot-Matrix



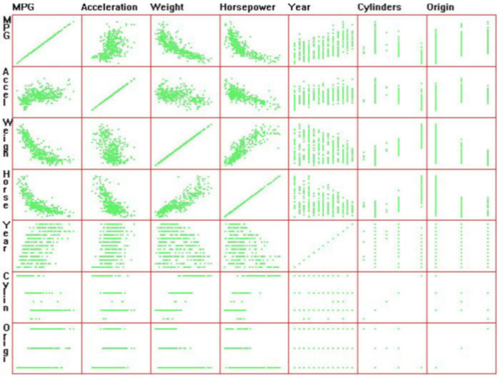

(a)




(b)

# Multi-Variate Daten: Scatterplot-Matrix

- Eigenschaften
  - Scatterplots auf der Diagonale
    - Gleiches Attribut für $x$- und $y$-Achse
    - Besser: Histogramm (Bar-Chart)
  - Oberes und unteres Dreieck der Matrix enthalten gespiegelte Scatterplots
- $\rightarrow \frac{m^2-m}{2}$ **verschiedene** binäre Korrelationen



(b)

# Multi-Variate Daten: Scatterplot-Matrix

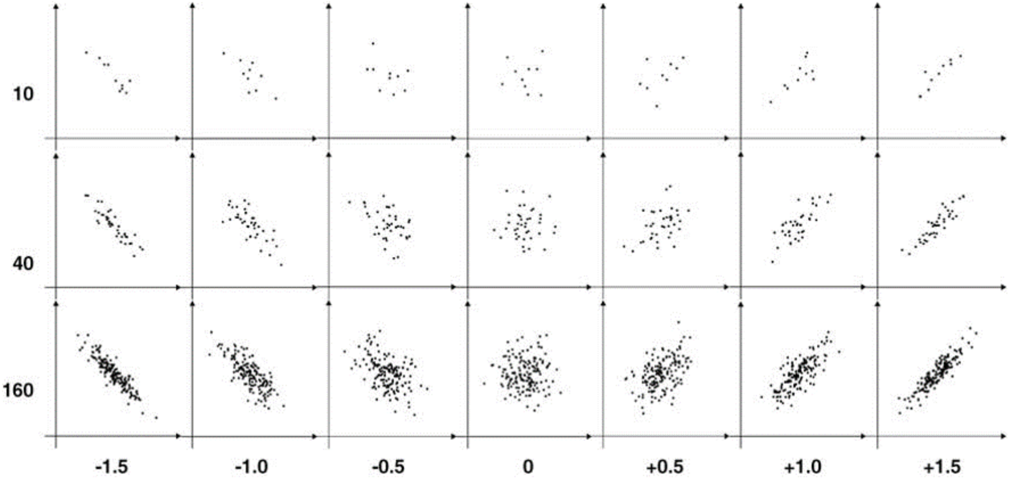



Abbildung: Scatterplot [LMvW2010]: Erkennung von Korrelationen

# Multi-Variate Daten: Hyperslice

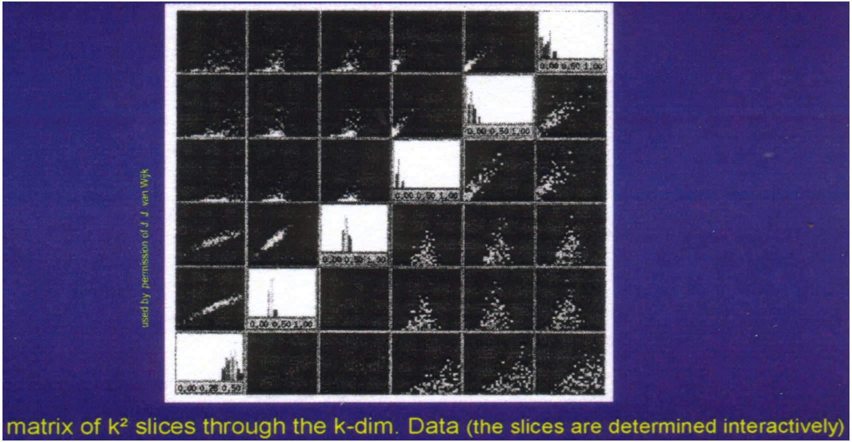



Abbildung: [vWvL1993],[AGK2002]

# Multi-Variate Daten: Prosection-Views

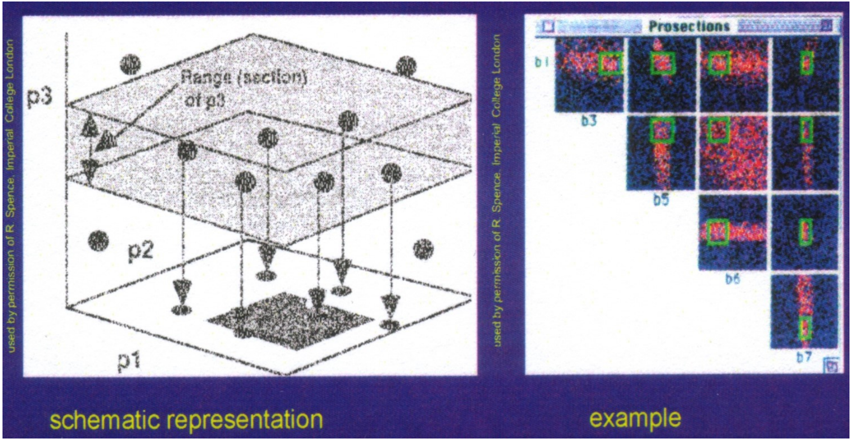



Abbildung: [FB1994], [SDTS1995], [AGK2002]

# Multi-Variate Daten

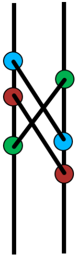
Achsenbasierte Ansätze

1. Scatterplot & Scatterplot Matrizen
2. Parallele Koordinaten
3. Star Plot
4. Flexible Linked Axes

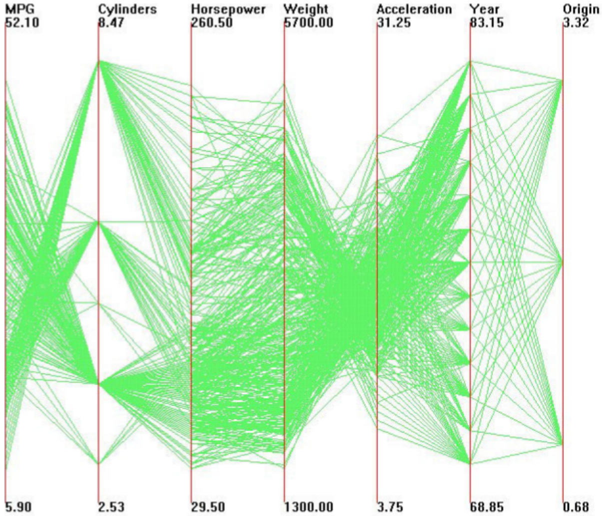
# Multi-Variate Daten: Parallele Koordinaten

- $m$ Attribute $\rightarrow$ $m$ Achsen
- Anordnung der Achsen: parallel
  - Scatterplot: orthogonal
- Jede Achse $j$ wird skaliert:
  - $\min_{1 \leq i \leq n}(x_{ij})$
  - $\max_{1 \leq i \leq n}(x_{ij})$
- Für jedes $i$ wird ein Polygonzug mit Kanten $(x_{i,j}, x_{i,j+1})$, $1 \leq j \leq m-1$ gezeichnet

- Beispiel
  - $n = 3, m = 2$
  - blau und rot: gleiche Richtung
  - grün: entgegengesetzt zu blau und rot



# Multi-Variate Daten: Parallele Koordinaten




(a)

# Multi-Variate Daten: Parallele Koordinaten

- Linienzüge zeigen lineare Abhängigkeiten der Daten
- Die Polygonzüge schneiden sich zwischen zwei Achsen in maximal einem Punkt
- Man kann Regeln für *k*-dimensionale Unterräume ableiten [Inselberg 1998]

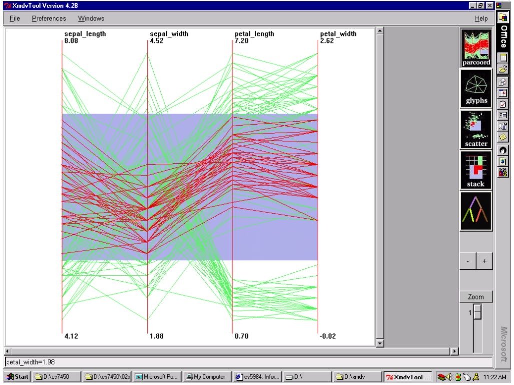



Abbildung: [http://davis.wpi.edu/ xmdv/]

# Multi-Variate Daten: Parallele Koordinaten

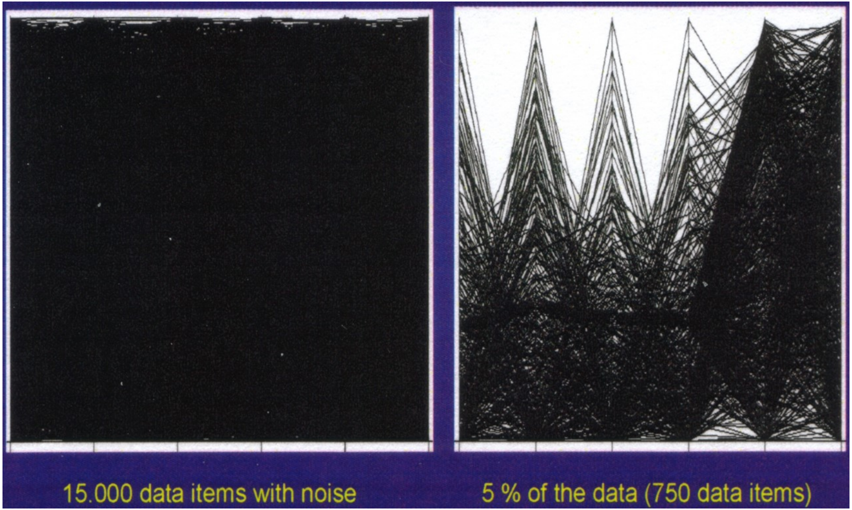



Abbildung: Leichtes Rauschen verursacht Probleme [AGK2002]

# Multi-Variate Daten: Parallele Koordinaten

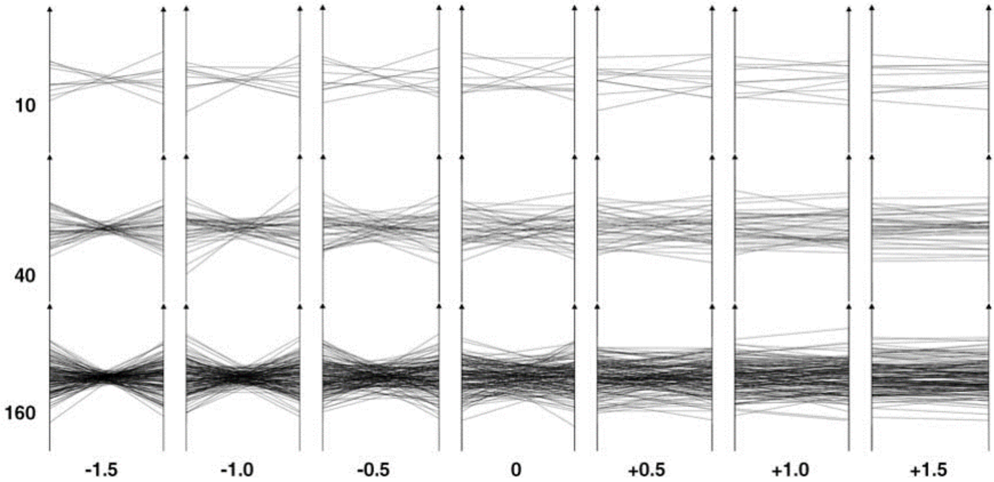



Abbildung: Parallele Koordinaten [LMvW2010]: Erkennung von Korrelationen

# Multi-Variate Daten: Parallele Koordinaten

- Eigenschaften
  - Anzeige von $m-1$ Korrelationen
  - Reihenfolge der Achsen bestimmt, welche Korrelationen sichtbar sind
  - [HYFC2014]
    - Positive Korrelationen und deren Stärke sind relativ schwer zu erkennen
    - Negative Korrelationen und deren Stärke sind relativ einfach zu erkennen
  - Mehr Überlappungen als bei Punkt-basierten Darstellungen
- Erweiterungen
  - Brushing (Parvis)
  - Transferfunktionen
  - Clusterbildung
  - Fokus und Kontext (Kapitel "Interaktion")

# Multi-Variate Daten: Parallele Koordinaten

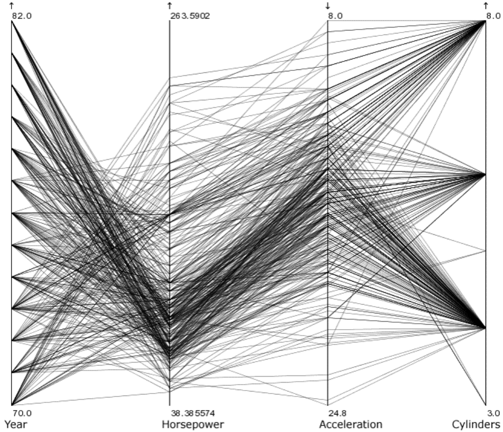



Abbildung: Parvis

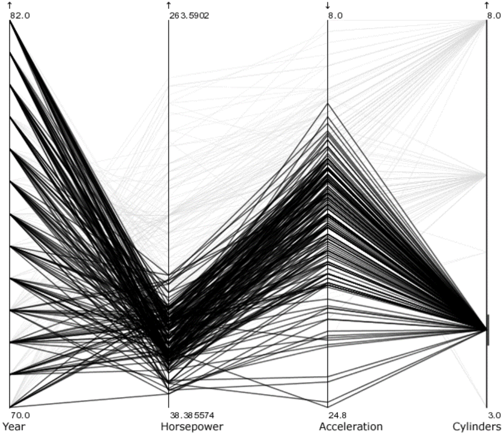



Abbildung: Parvis: Brushing

# Multi-Variate Daten: Parallele Koordinaten

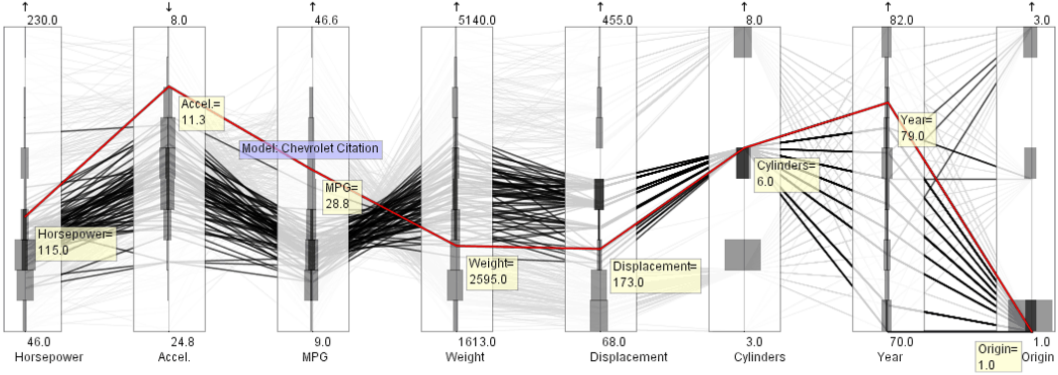



Abbildung: Parvis: Brushing + Histogram

# Multi-Variate Daten: Parallele Koordinaten

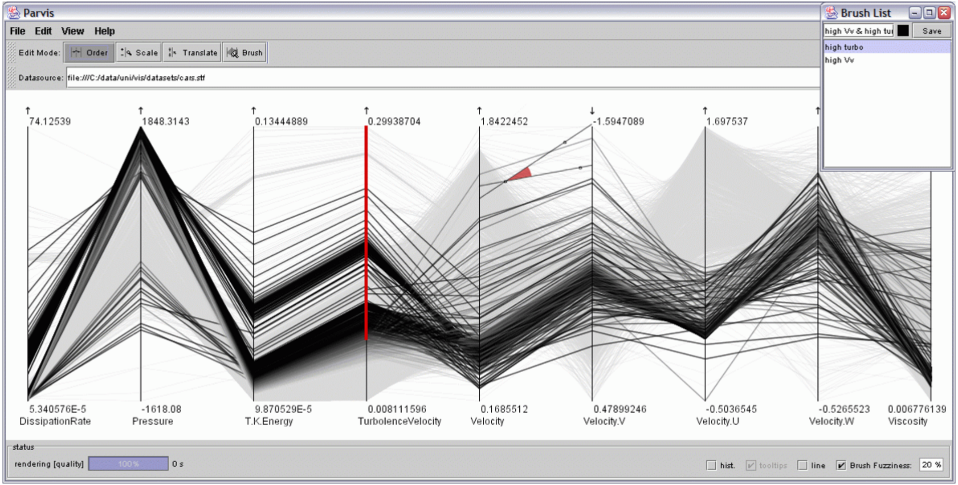



Abbildung: Parvis: Auswahl über Achsenabschnitte und Winkel (in rot) → Fokus

# Multi-Variate Daten: Parallele Koordinaten

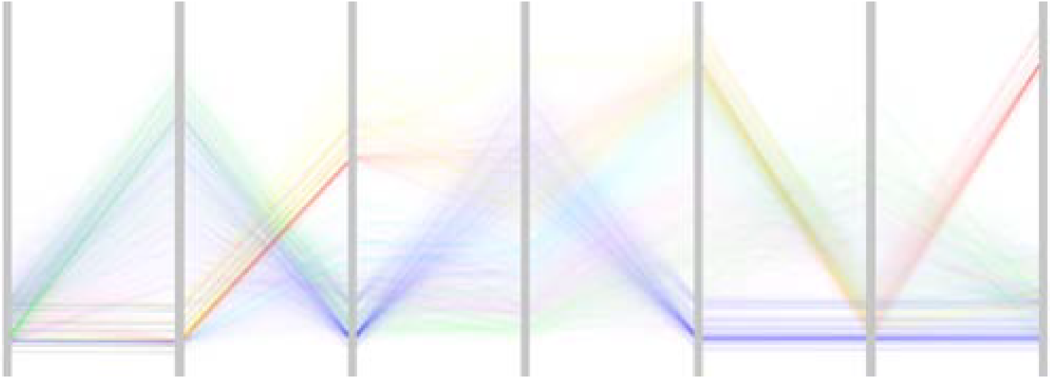


Abbildung: Lineare Transferfunktion angewendet auf eine hochauflösende Textur um Überdeckungen zu reduzieren und einen Überblick über die Daten zu geben. [Johan2005]

# Multi-Variate Daten: Parallele Koordinaten

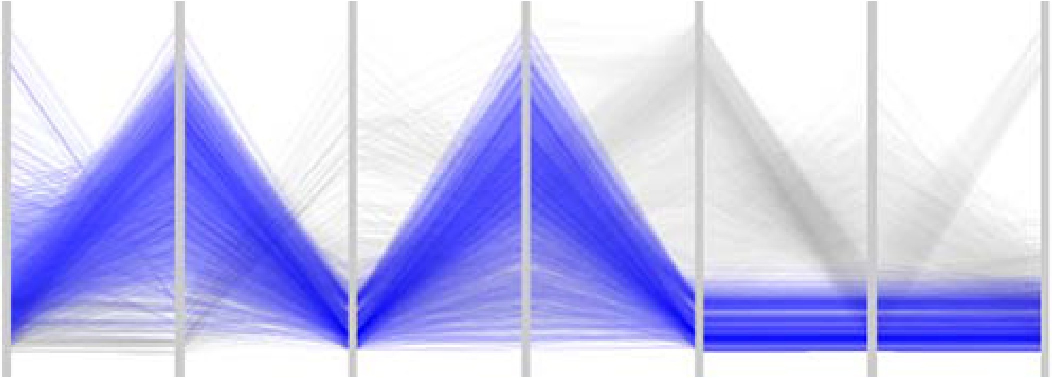


Abbildung: Logarithmische Transferfunktion angewandt auf einen ausgewählten Cluster: Die Struktur bleibt erhalten und Regionen geringer Dichte werden hervorgehoben. [Johan2005]

# Multi-Variate Daten: Parallele Koordinaten

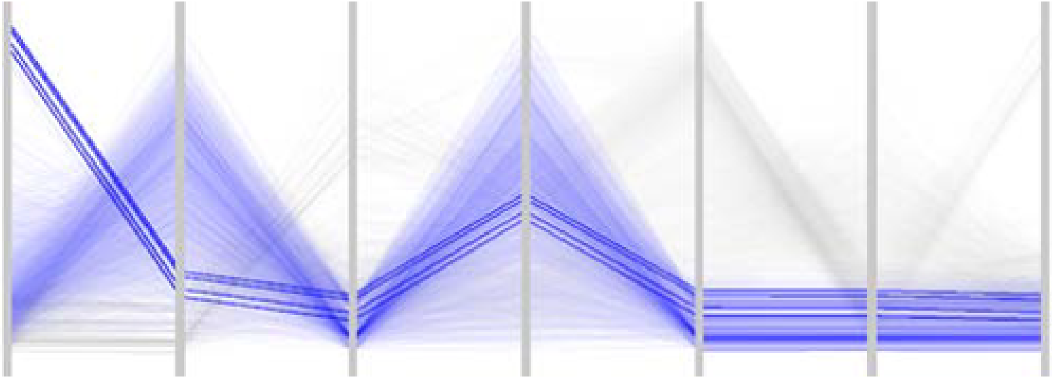


Abbildung: Transferfunktion unter Verwendung der Quadratwurzel: Ausreißer von lokalen Clustern werden hervorgehoben. [Johan2005]

# Multi-Variate Daten: Parallele Koordinaten

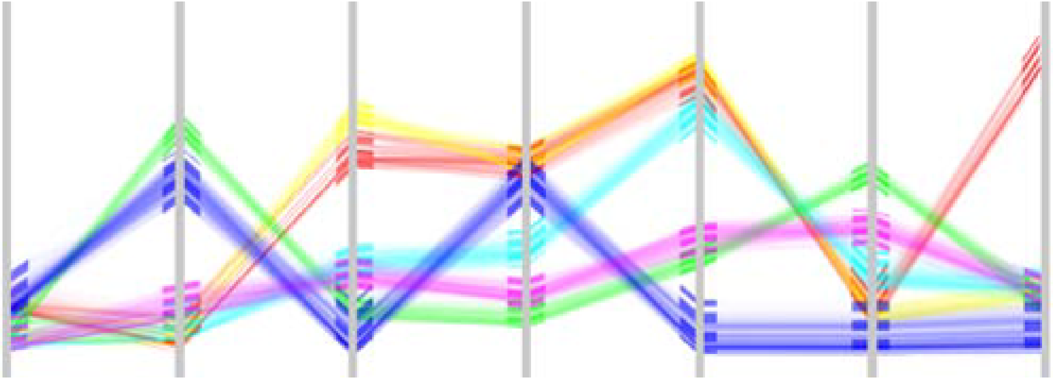
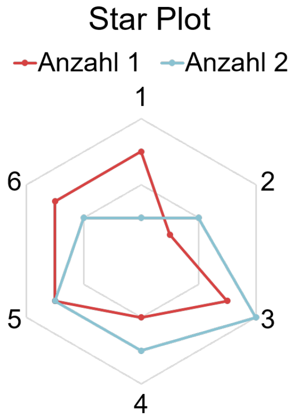


Abbildung: Komplementäre Ansicht der Cluster: uniforme Bandstruktur. [Johan2005]

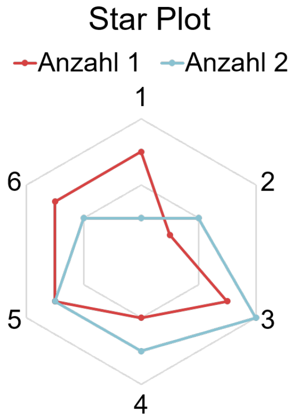
# Multi-Variate Daten: Star-Plot

- $m$ Attribute $\rightarrow$ $m$ Achsen
- Anordnung der Achsen: Stern-förmig
- Verbindung von $x_{i,j}$ mit $x_{i,j+1}$, $0 \leq j \leq m-1$
- Zusätzliche Verbindung von $x_{i,m}$ mit $x_{i,1}$
- Anzeige von $m$ Korrelationen
- Ergibt geschlossenen Polygonzug (Polygon)
- Meist zum qualitativen Vergleich der Form



# Multi-Variate Daten: Star-Plot

- Andere Bezeichnungen
  - Radar-Chart
  - Kiviat-Diagramm



# Multi-Variate Daten: Star-Plot Varianten

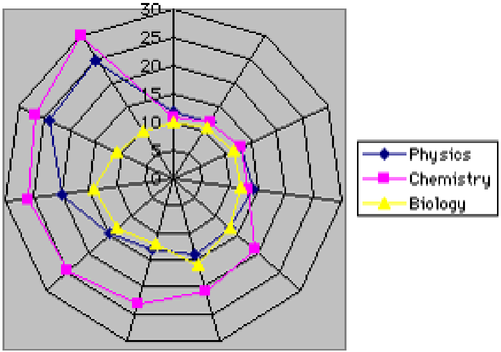



Abbildung: Ein Stern für alle Datenpunkte

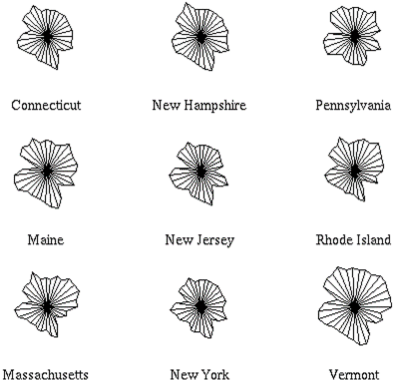
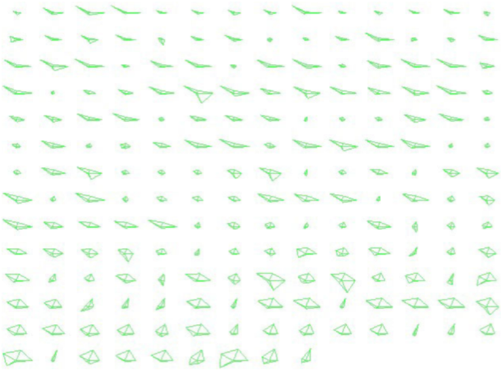


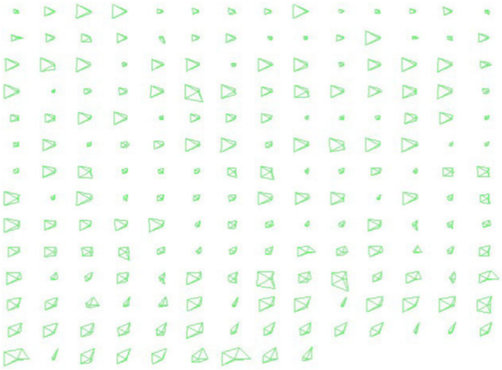

Abbildung: Ein Stern pro Datenpunkt

[http://seamonkey.ed.asu.edu/~behrens/asu/reports/compre/comp1.html]

# Multi-Variate Daten: Star-Plot



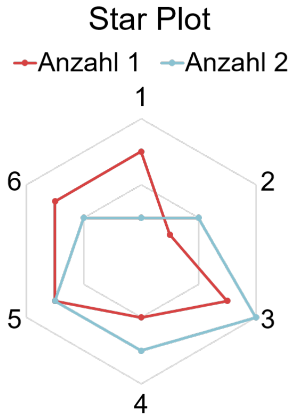
(a)



(b)

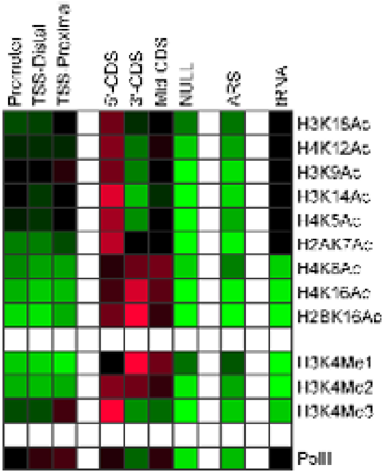
# Multi-Variate Daten: Star-Plot

- Eigenschaften
  - siehe auch: Parallele Koordinaten
  - Unterschiedliche Qualität der Achsen durch unterschiedliche Richtungen
  - Abstand zwischen zwei Achsen
    - Klein nahe dem Zentrum
    - Groß nahe dem Rand



# Multi-Variate Daten: Heatmap

- Darstellung von Korrelationen zwischen zwei Attributen in einer Matrix
- Stärke der Korrelation im Intervall $[-1;1]$
- Mapping
  - 0: neutral $\rightarrow$ schwarz
  - $< 0$: rot
  - $> 0$: grün
- Absoluter Betrag des Wertes: Helligkeit
- Mapping ungünstig bei rot-grün Sehschwäche



# Multi-Variate Daten

- Einige Techniken sind besonders geeignet für
  - kategorische Attribute
  - ordinale Attribute
- Mischungen möglich mit
  - kontinuierlichen Attributen
- Ordinale Attribute
  - Führen mit einer kleinen Wertmenge zu auffälligen Clustern
    - Diese springen dem Anwender ins Auge
    - Überbetonen damit diese Attribute

# Literatur


[Spence2001] R. Spence.
Information Visualization.
Addison-Wesley, Reading, MA, USA, 2001.

[WGK2011] M. Ward, G. Grinstein, D. Keim.
Interactive Data Visualization: foundations, techniques, and applications.
A K Peters. Ltd, 2011.

[AGK2002] M. Ankerst, G. Grinstein, D. Keim.
Visual Data Mining.
Tutorial at KDD 2002.

# Literatur


[Cleveland] W. S. Cleveland.
Visualizing Data.
AT&T Bell Labortories, Murray Hill, NJ, revised edition.

[vWvL1993] Jark van Wijk, Robert van Liere.
Hyperslice.
Proc. IEEE Visualization, 1993.

[FB1994] George W. Furnas, Andreas Buja.
Prosection Views: Dimensional Inference through Sections and Projections.
Journal of Computational and Graphical Statistics,
Vol. 3, No. 4, 1994, pp. 323–353.

# Literatur


[SDTS1995] H. Su, H. Dawkes, L. Tweedie, R. Spence.
An Interactive Visualization Tool for Tolerance Design.
Technical Report, Imperial College, London, 1995.

[AC1991] Alpen und Carten.
Hyperbox.
Proc. IEEE Visualization, pp. 133–139, 1991.

[HYFC2014] Lane Harrison, Fumeng Yang, Steven Franconeri, Remco Chang.
Ranking Visualizations of Correlation Using Weber's Law.
IEEE TVCG Vol. 20(12), 2014.

# Literatur

[Joha2005] Jimmy Johansson.
Revealing Structure within Clustered Parallel Coordinates Displays.
InfoVis, 2005.