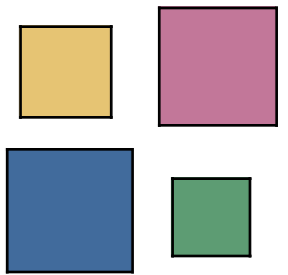


## Visualisierung hochdim. kategoriemer Daten

- Multivariate Daten mit  $k$  Variablen mit je  $n_1, \dots, n_k$  Stufen d.h. wir betrachten nur  $\prod_{i=1}^k n_i$  Werte.

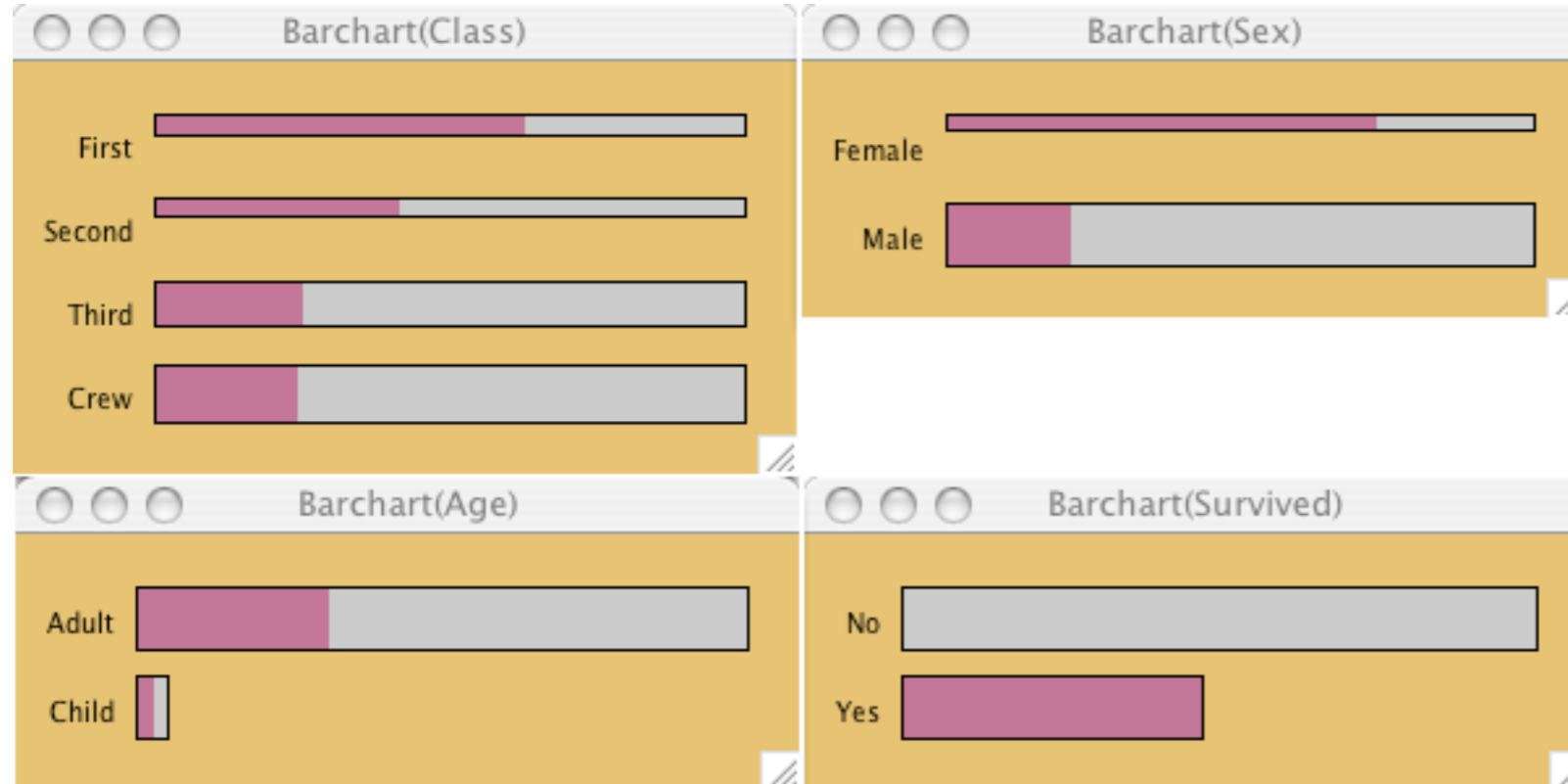
Beispiel Titanic Daten: 32 Werte aus 2201 Passagieren

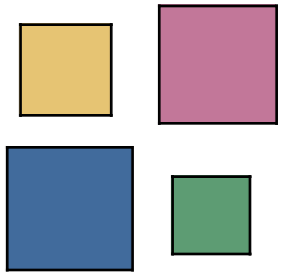
- Klasse: 1., 2., 3., Crew  $\rightarrow$  4
  - Alter: Erwachsene, Kinder  $\rightarrow$  2
  - Geschlecht: Männer, Frauen  $\rightarrow$  2
  - Überlebt: Ja, Nein  $\rightarrow$  2
- Linked Highlighting kann nur bedingte Verteilungen liefern  $\Rightarrow$  2-dimensional.
  - Mögliche Lösungen:
    - Dimensionsreduktion  $\Rightarrow$  “klassische Visualisierung”
    - Visualisierung von Parametern
    - Visualisierung der multivariaten Daten



## Linking

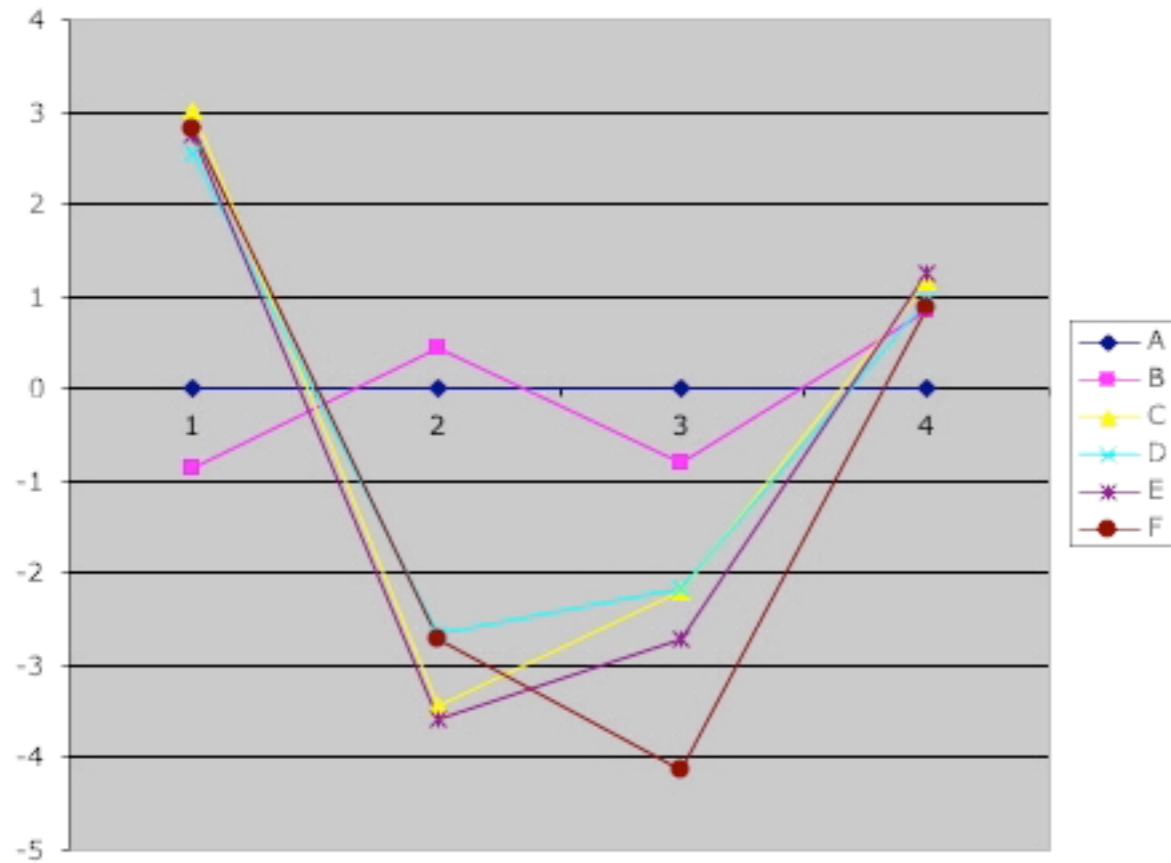
- Jeder Barchart zeigt die 1-dimensionale Projektion auf diese Variable
- Linking zeigt die bedingte Verteilung dieser Projektion gegeben die aktuell selektierte Gruppe
- Beispiel Titanic: bedingte Verteilungen für die Überlebenden



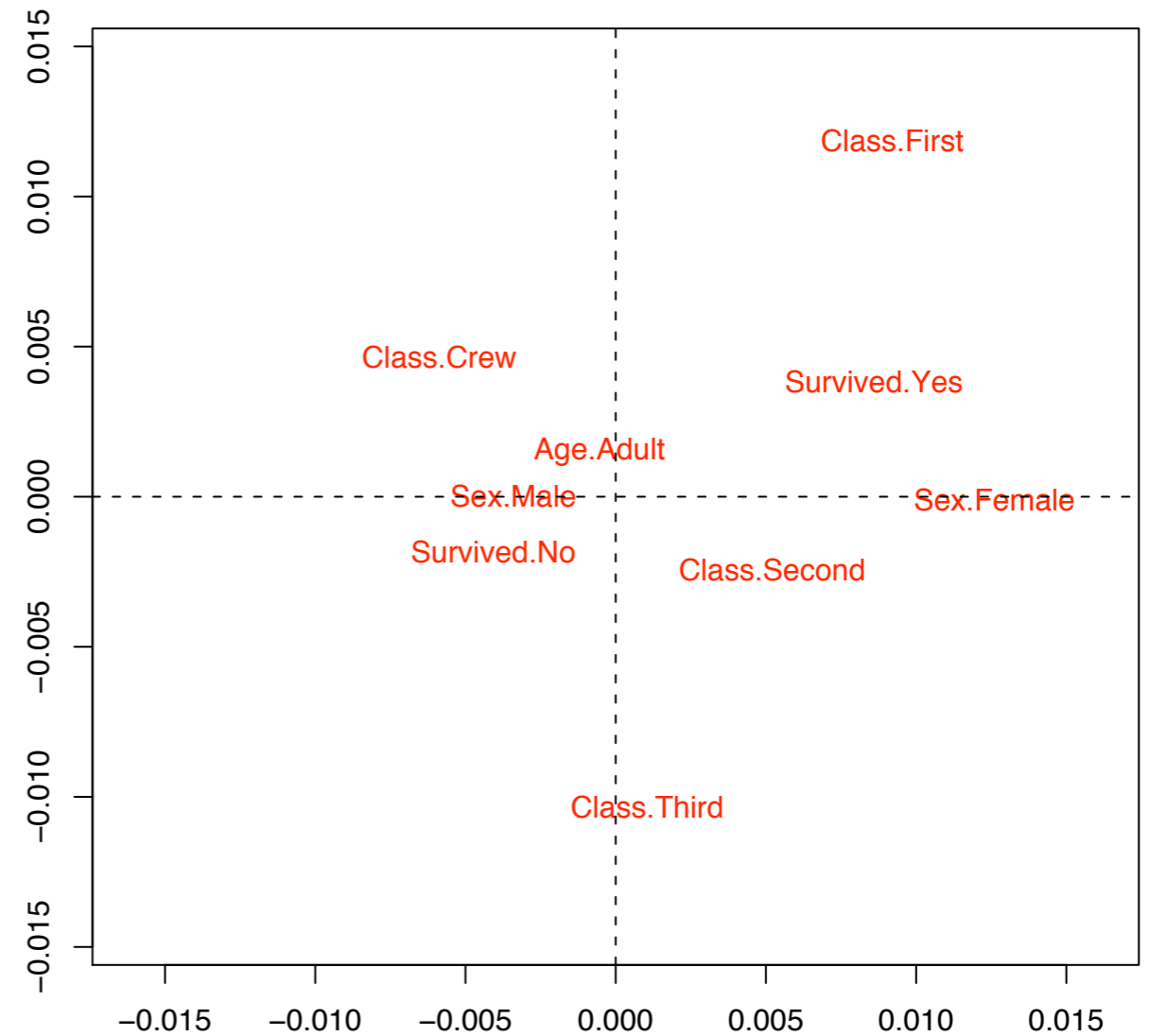


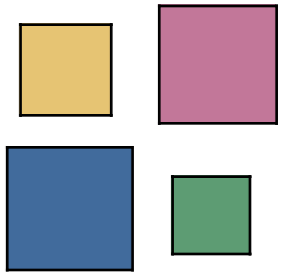
# Methoden zur Dimensionsreduktion

- Loglineare Modelle  
(Visualisierung der Parameter)



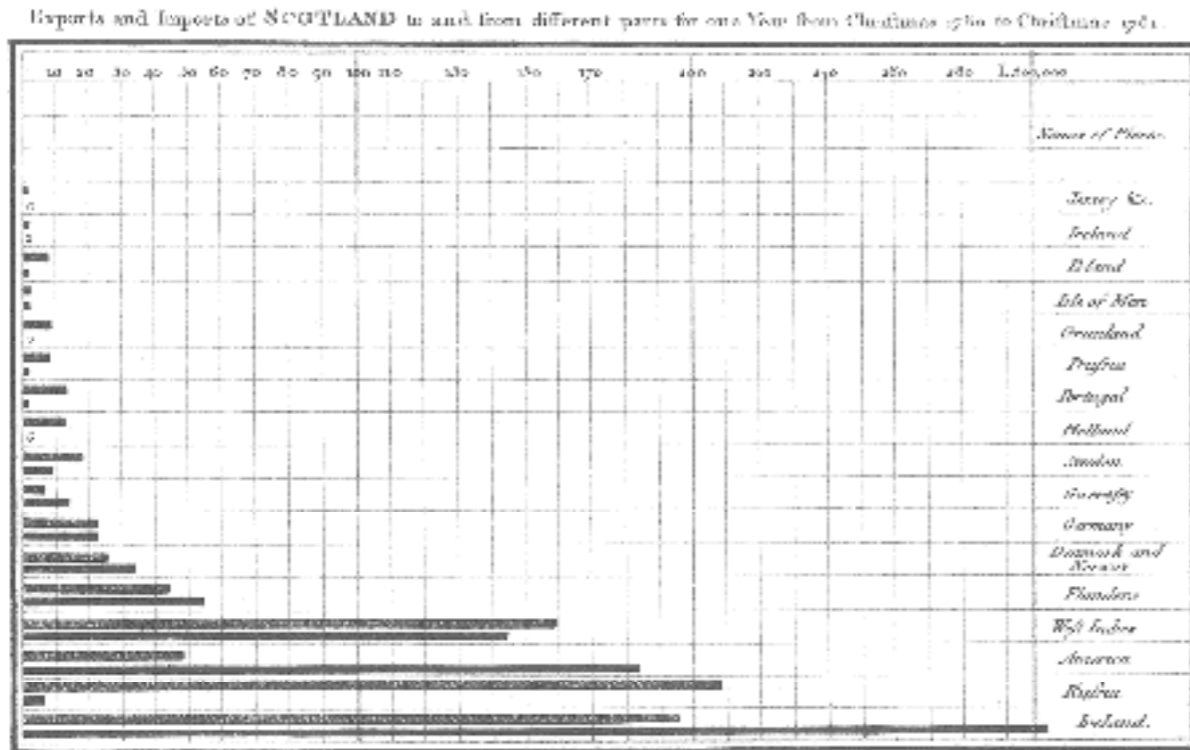
- Korrespondenz Analyse





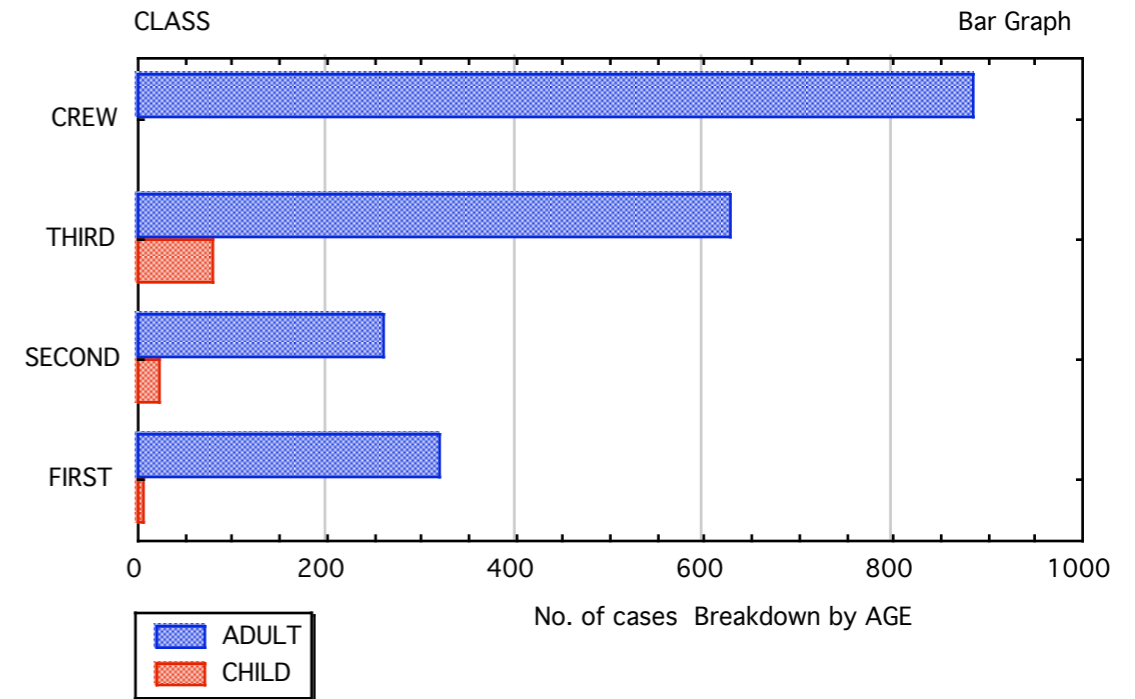
# Visualisierung: Historie

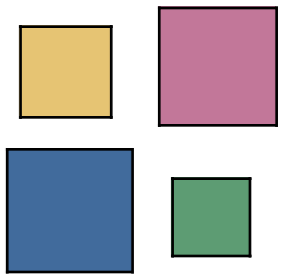
- William Playfair (1786)



The Upright divisions are Ten Thousand Pounds each The Black Lines are Exports the Red Lines Imports  
 Printed in the British Museum by W. Phillips  
 Richard D. Smith, London

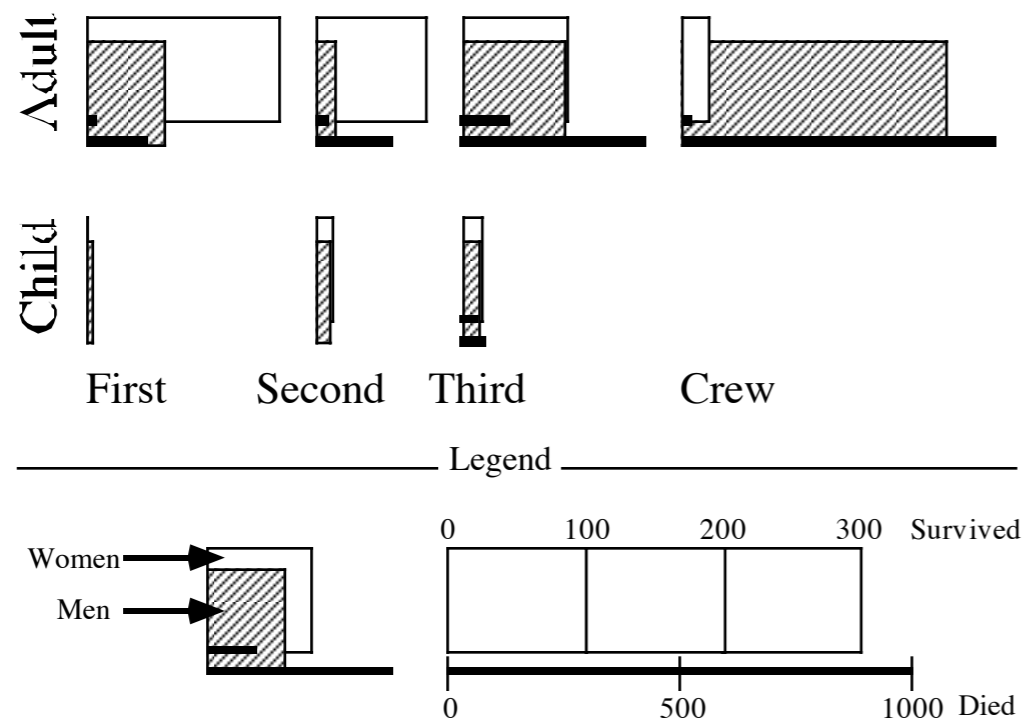
- Statistica (1996)



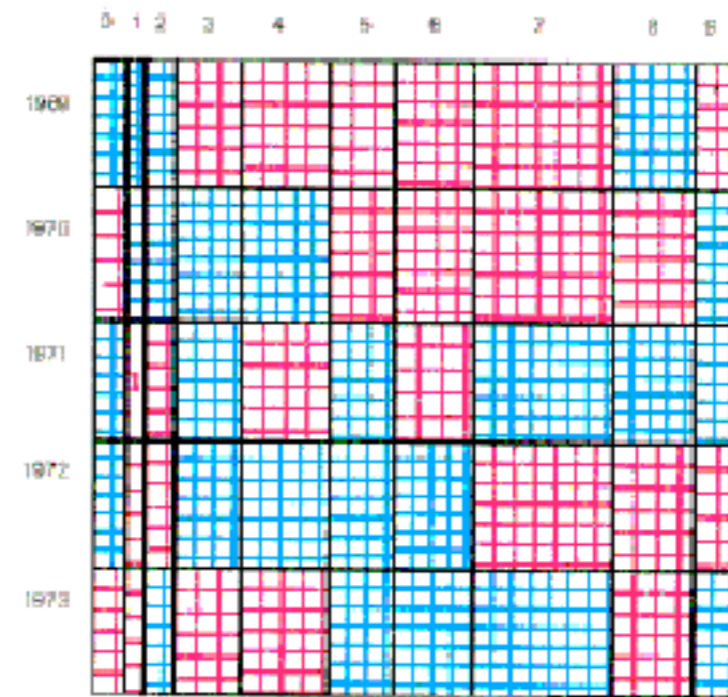


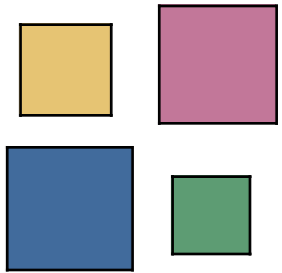
# Historie (cont.)

- Jaques Bertin (1967)



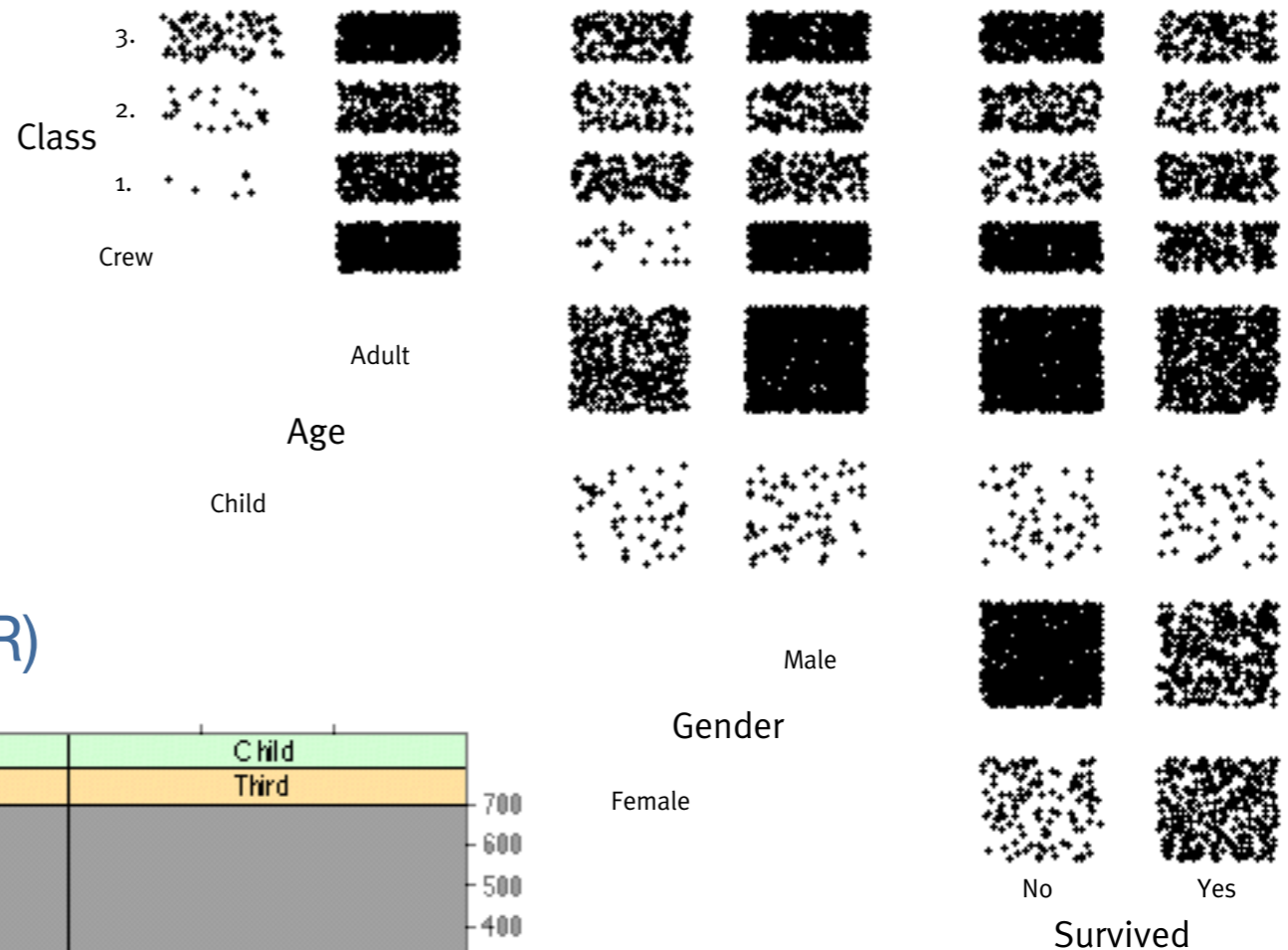
- Riedwyl & Schuepbach (1994)



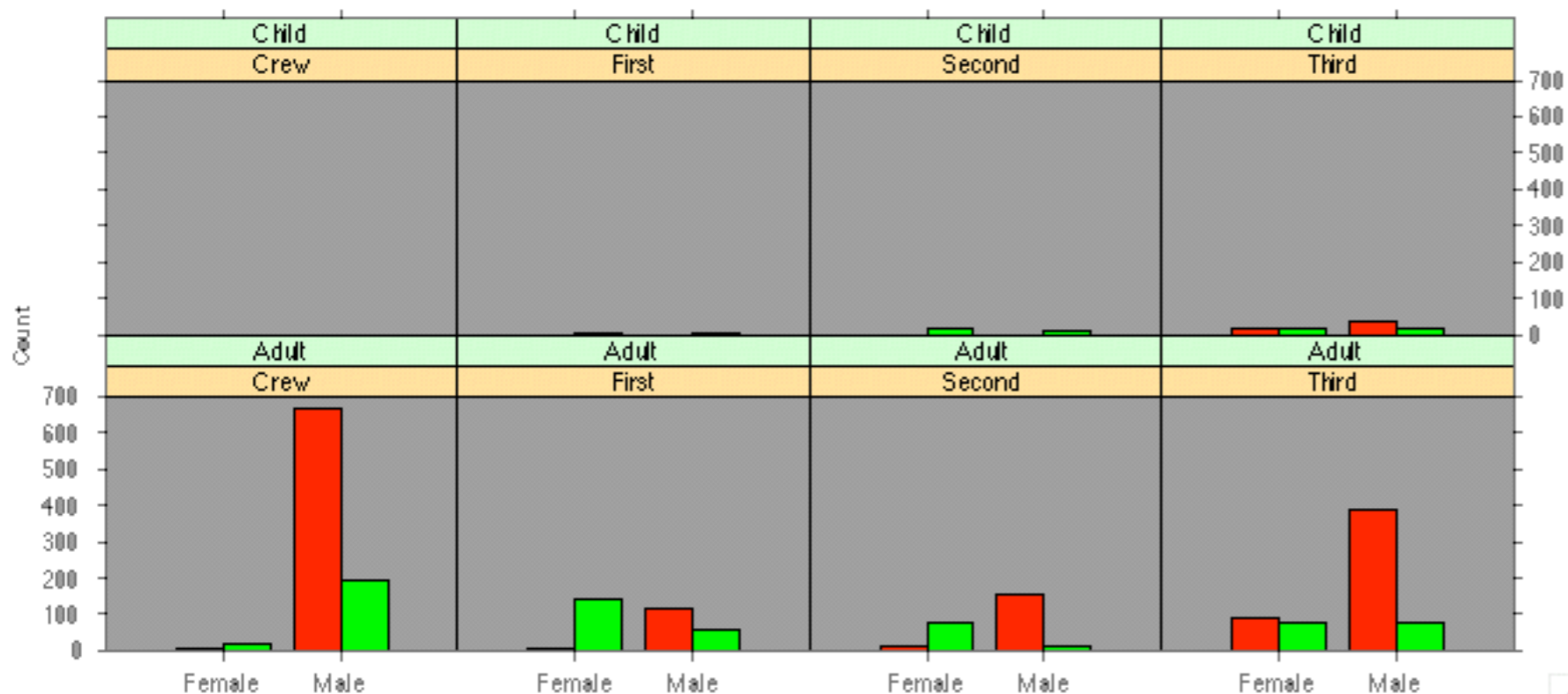


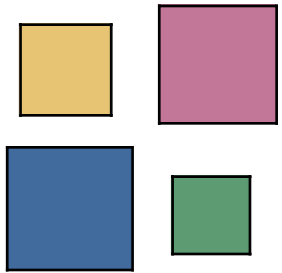
# Historie (cont.)

- Nagel & Ostermann (1992)

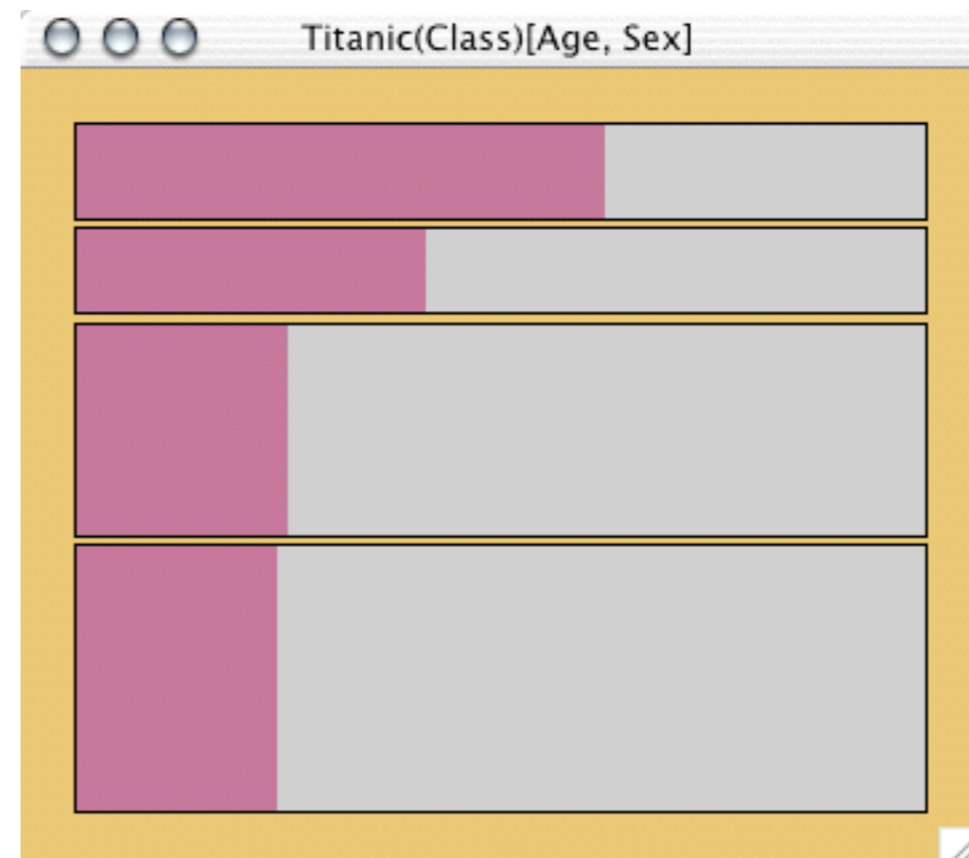
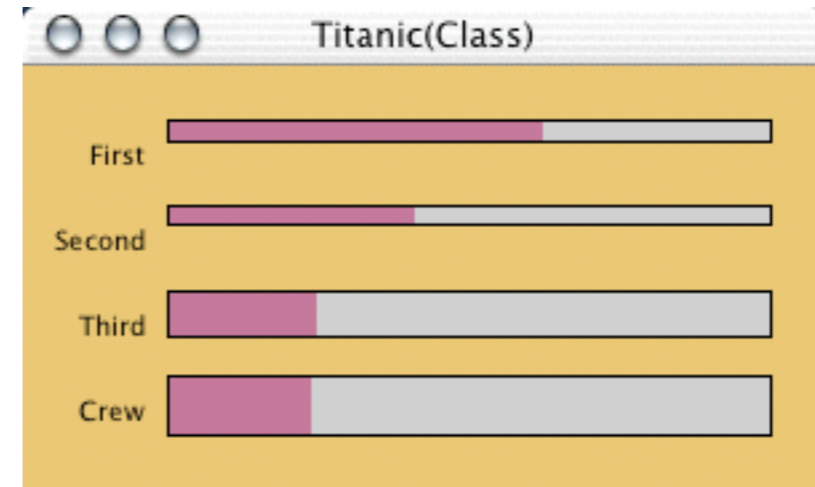
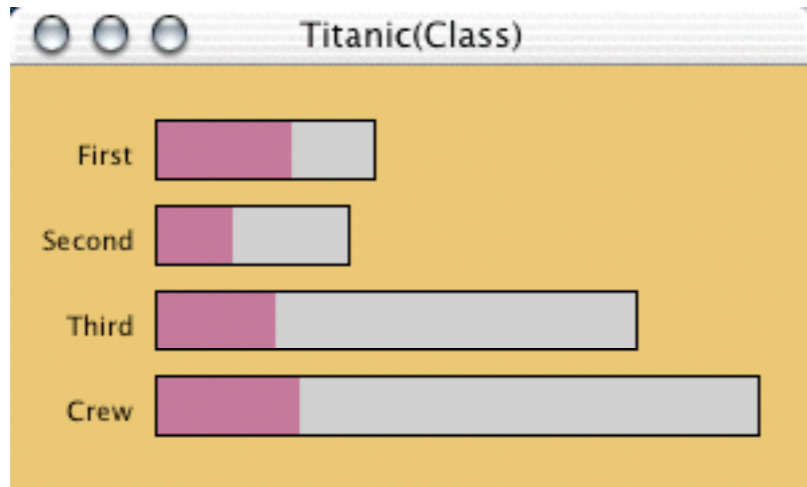


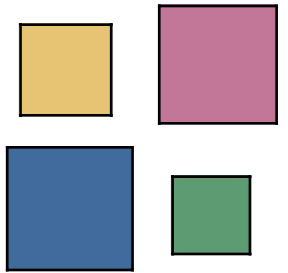
- W. Cleveland (1983)  
(Trellis Plots, aka Lattice Graphics in R)





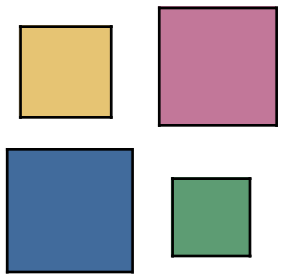
# Mosaic Plots: Konstruktion





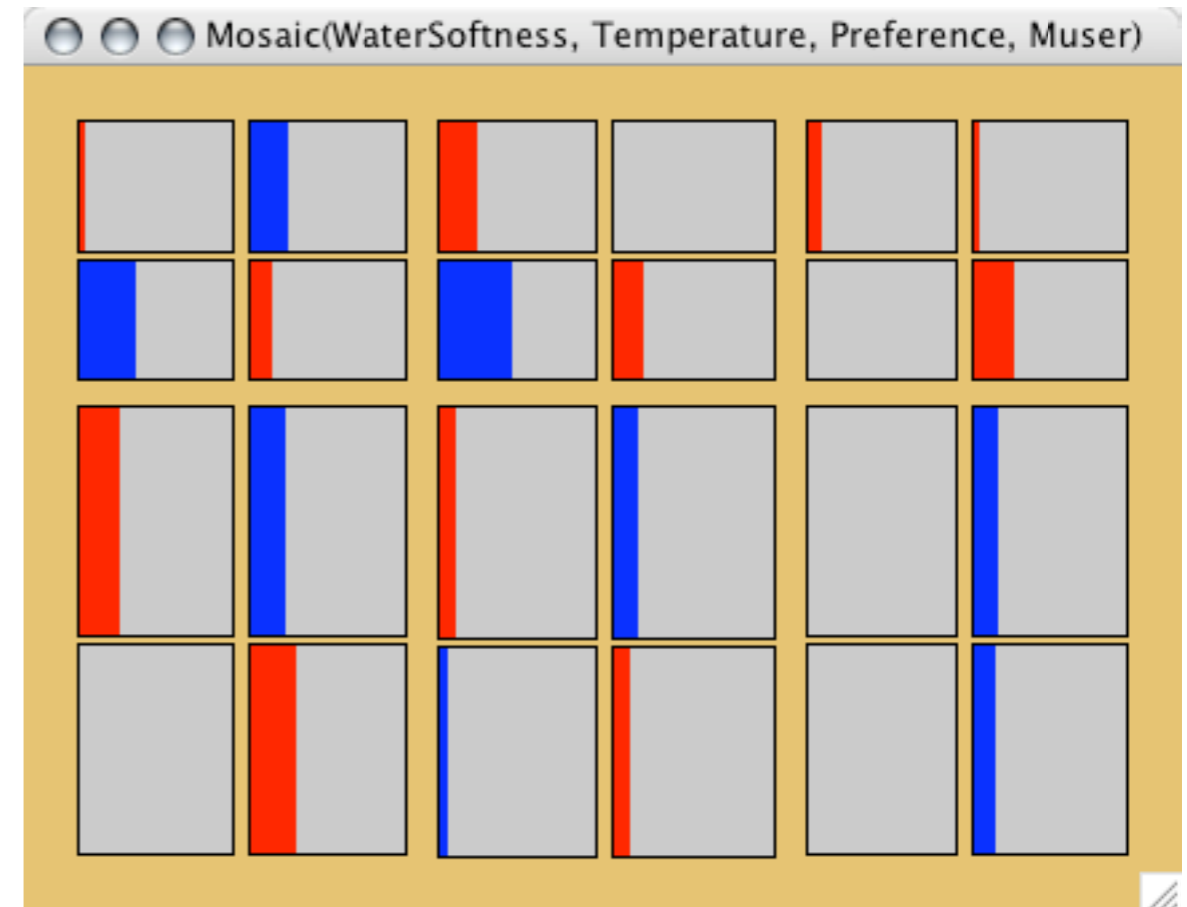
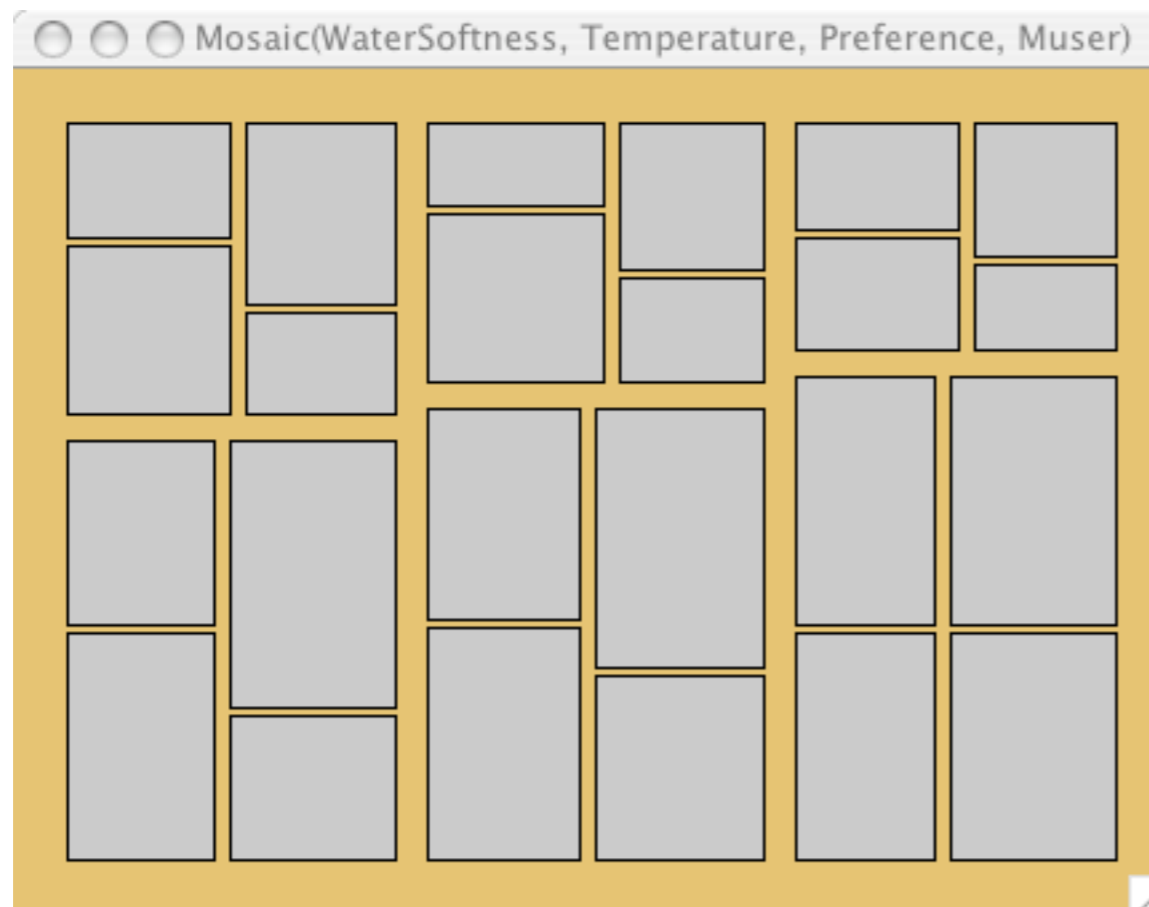
## Mosaic Plots: Eigenschaften

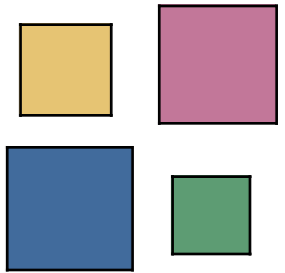
- Die Fläche der einzelnen Zellen ist proportional zur Anzahl der Fälle in dieser Zelle (analog zu Barchart und Histogramm)
- Die Zwischenräume sind zusätzlicher Platz
- Die Definition ist strikt rekursiv
  - ⇒ Reihenfolge ist wichtig
  - ⇒ Konditionierte Darstellung
- Umsortieren von Variablen und Stufen der Faktoren ist daher wichtig!
- Für eine einfache Interpretation ist eine Top-Down Entwicklung sinnvoll.
- Interaktionsstrukturen sind leicht zu erkennen (enge Verbindung zu loglinearen Modellen)



## Mosaic Plots und loglineare Modelle

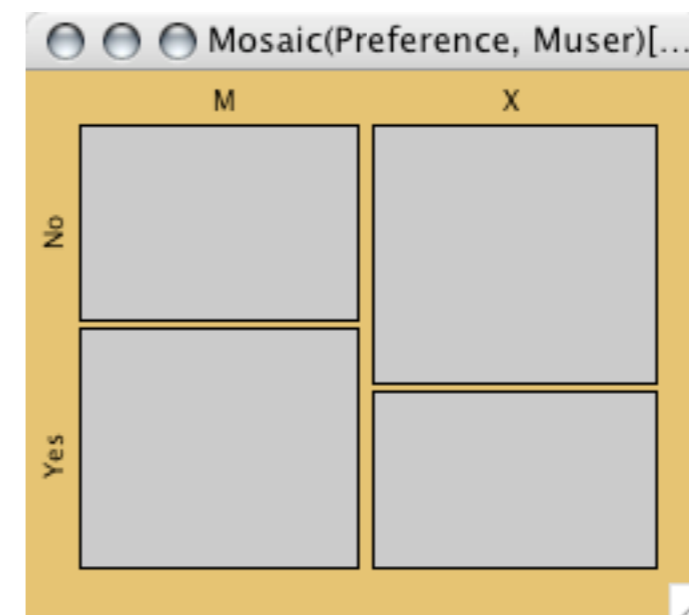
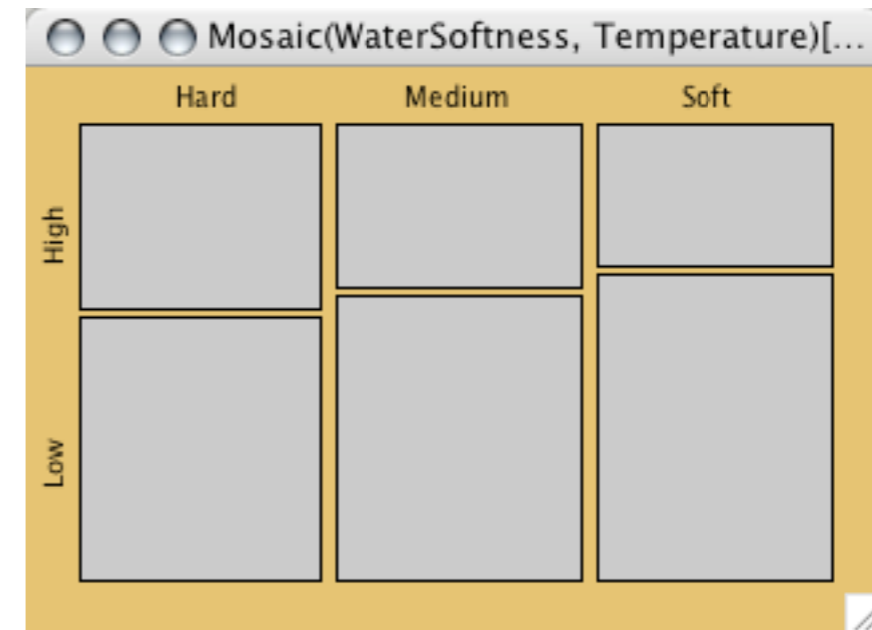
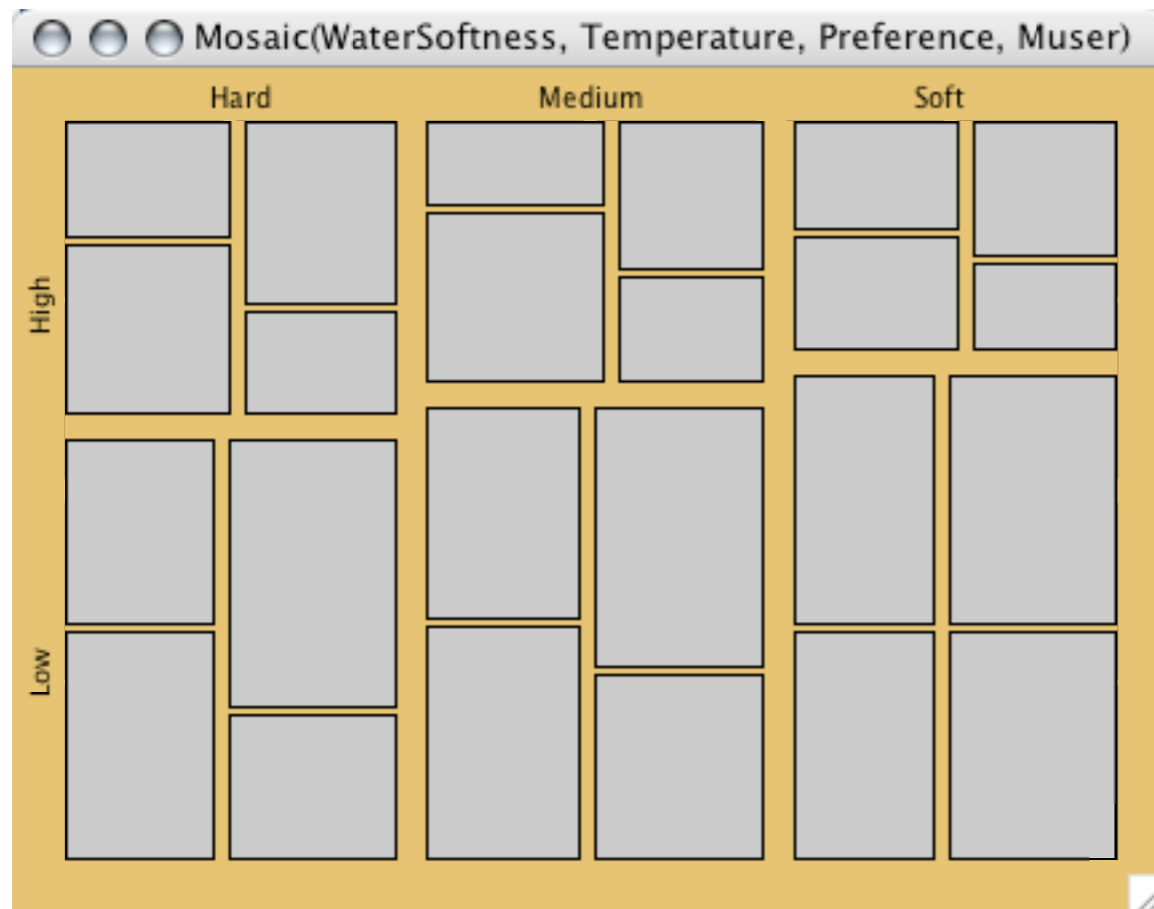
- Um ein loglineares Modell in einem Mosaic Plot dazustellen können anstatt der beobachteten Werte  $o_{ijk\dots}$  die erwarteten Werte  $e_{ijk\dots}$  dieses loglinearen Modells geplottet werden.
- Einfachster Fall: totale Unabhängigkeit
- **Beispiel: Detergent Daten**

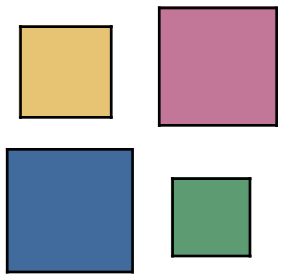




## Mehr zu loglinearen Modellen

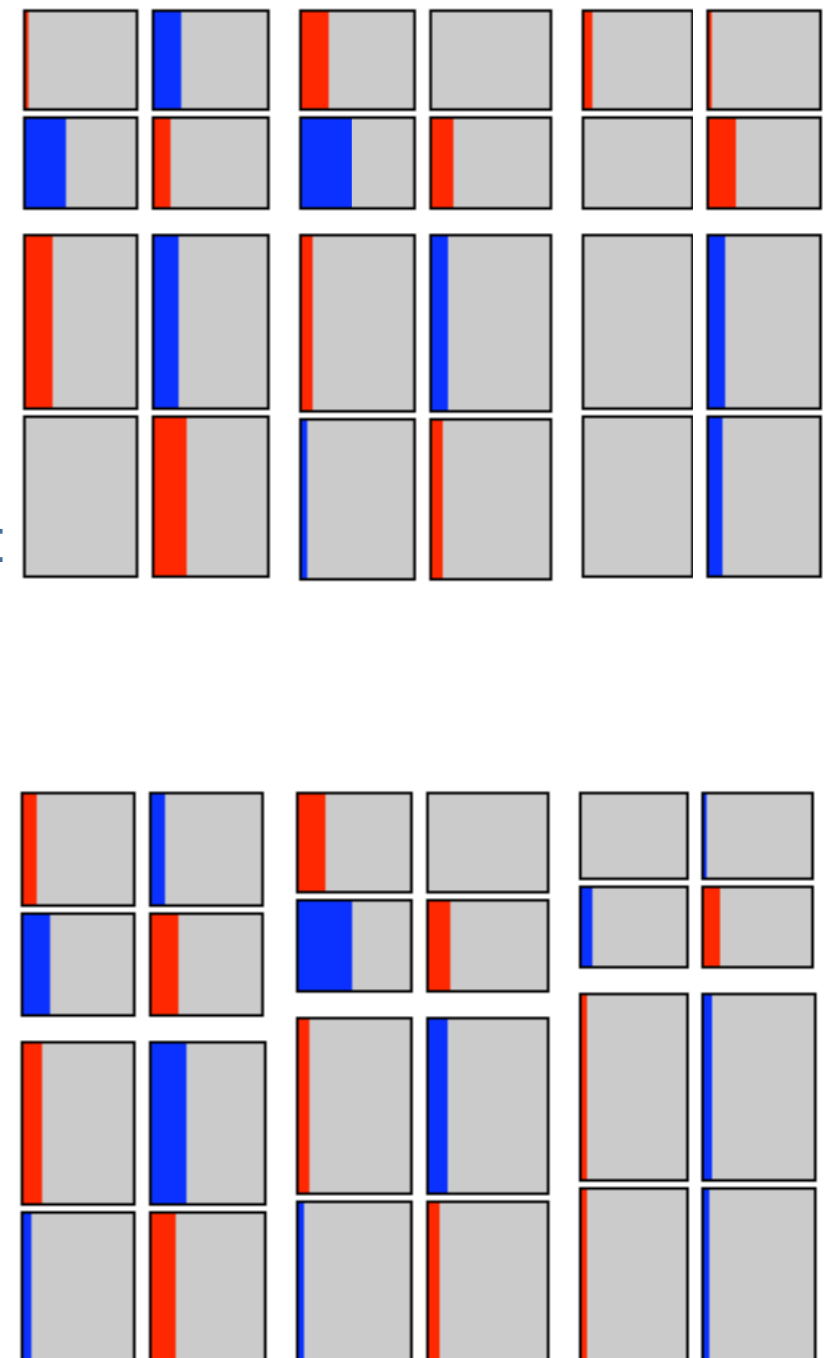
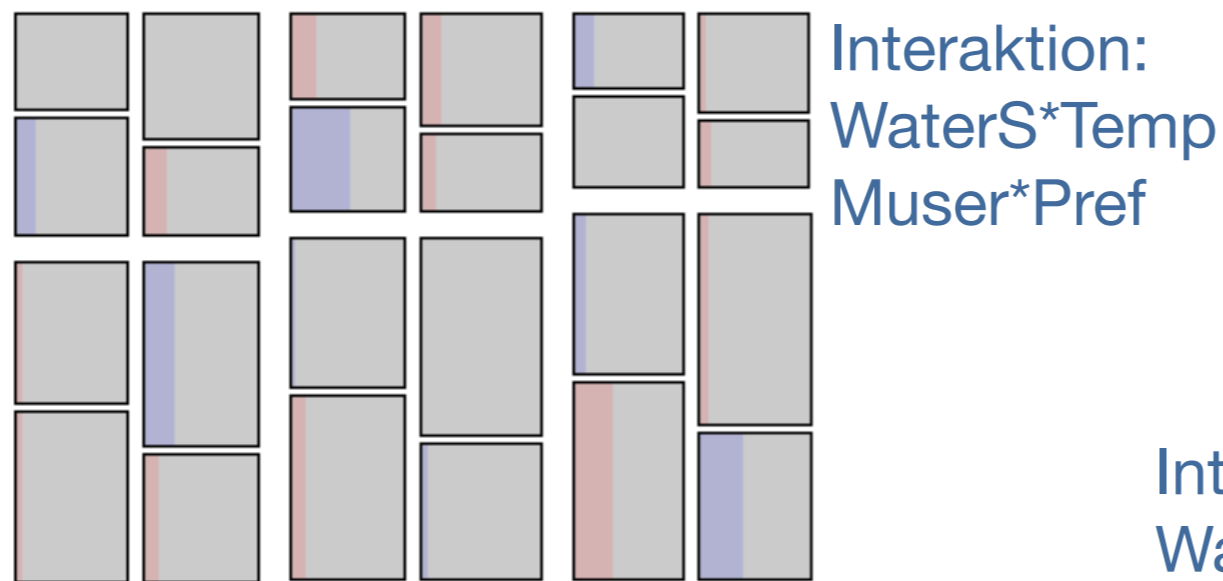
- Welche Interaktionen sind auffällig?

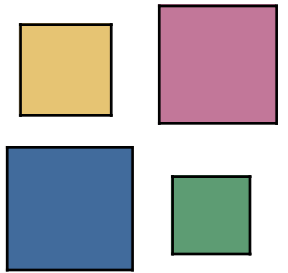




# Detergent (cont.)

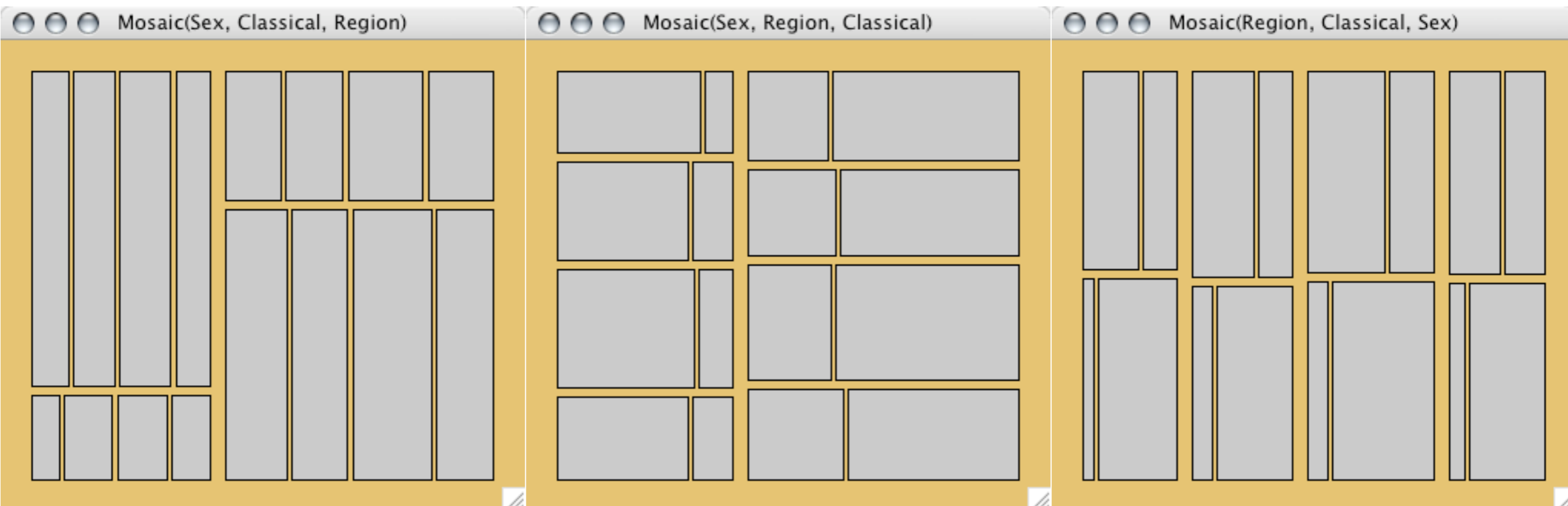
- Hinzufügen der Interaktionen

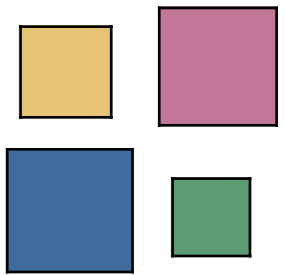




## Mehr zu Modellen

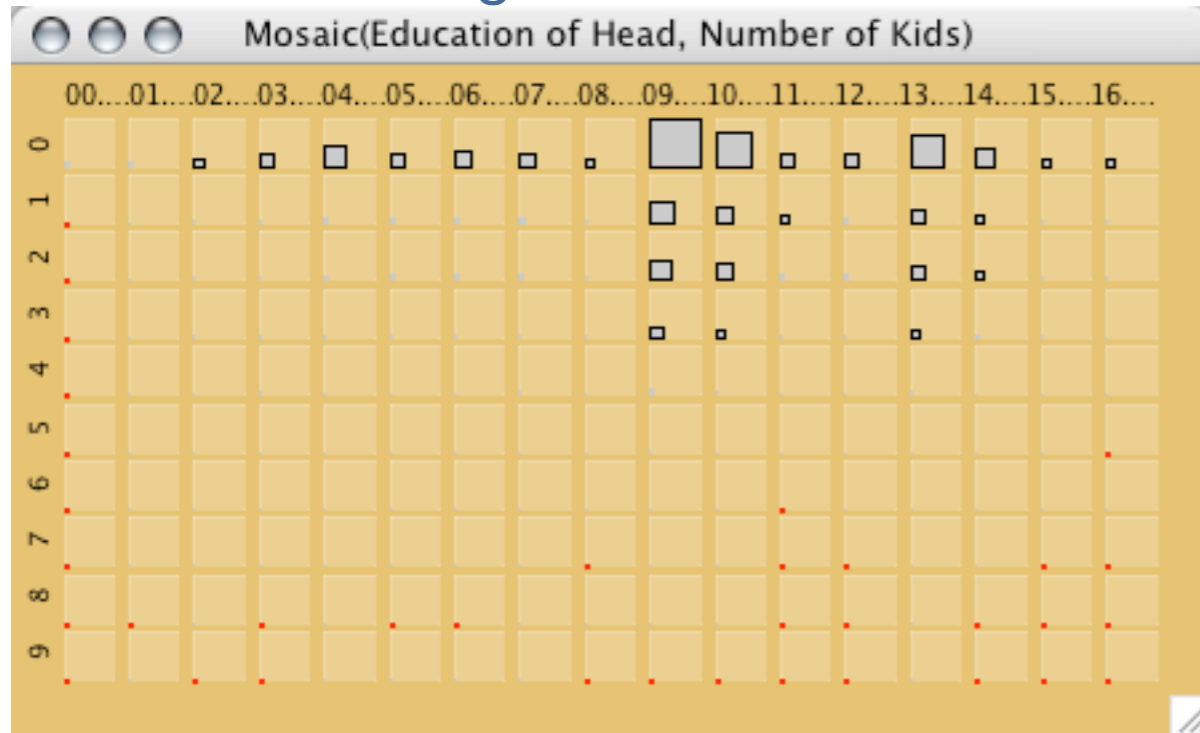
- Partielle Unabhängigkeit bei 3 Variablen:
  - d.h.  $\pi_{ijk} = \pi_{i++}\pi_{+jk} \quad \forall i, j, k$  bzw.  $\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$
  - es existiert nur eine Interaktion zwischen 2 Variablen,
  - d.h. die anderen 2 Paare sind unabhängig
- Je nach Reihenfolge ist die Abhängigkeitsstruktur besser oder schlechter zu erkennen  $\Rightarrow$  umordnen ist wichtig!



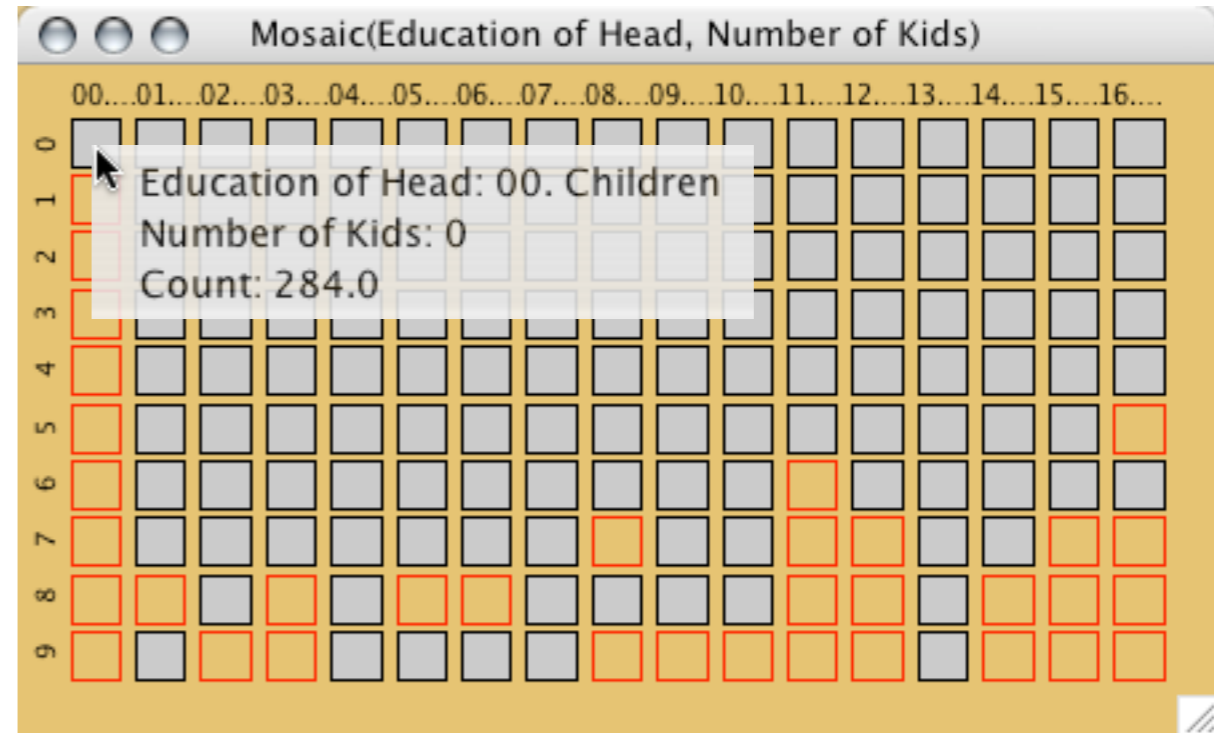


# Mosaic Plot Variationen

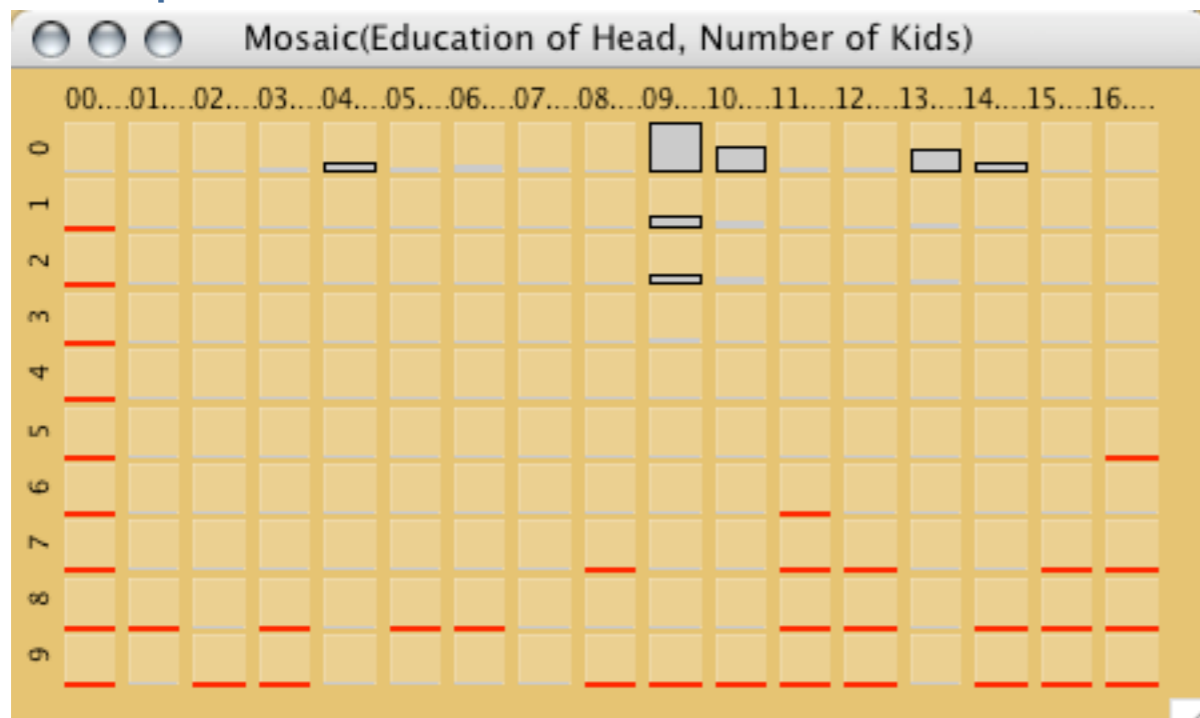
## Fluctuations Diagramm



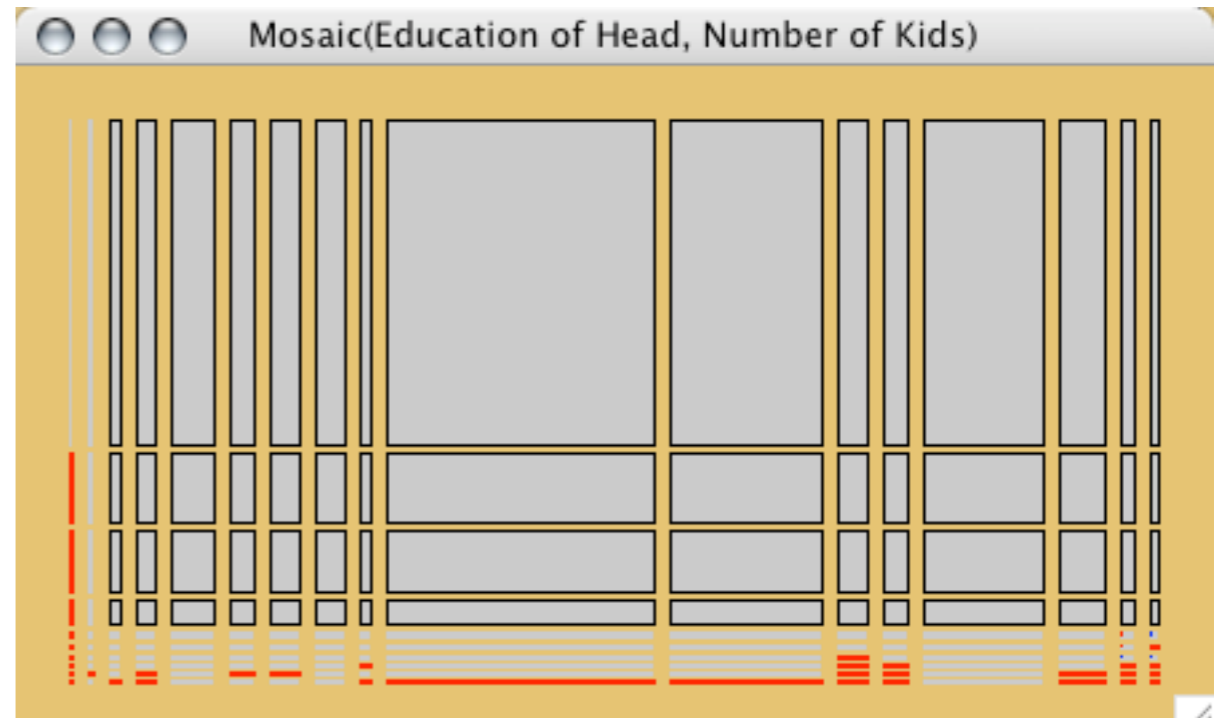
## Same Bin Size

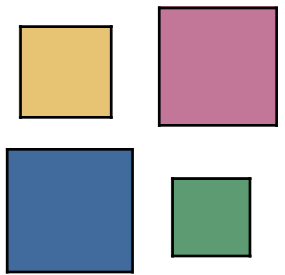


## Multiple Barchart



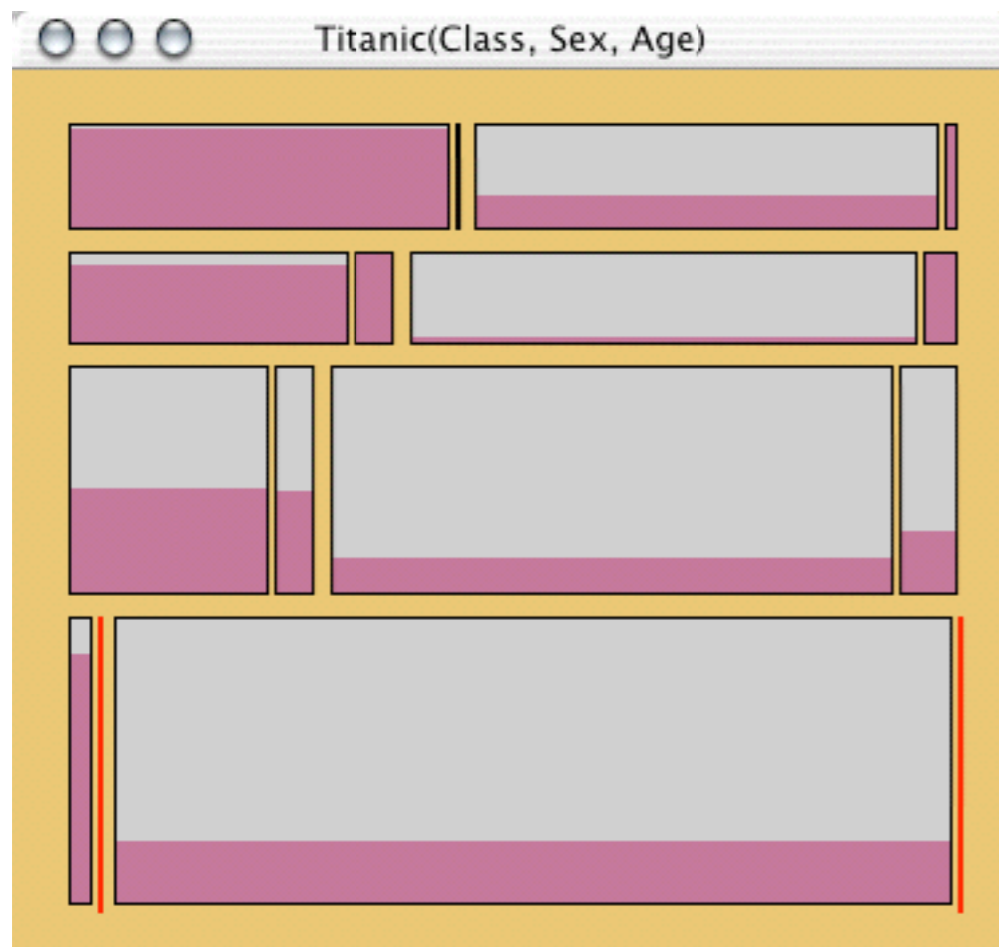
## Modelle ...





## Änderung der Reihenfolge/Orientierung

- Standard Aufteilung:  $xyxyx\dots$
- Je nach Daten und/oder Fragestellung können jedoch auch andere Aufteilungen sinnvoll sein, z.B.



yxx

