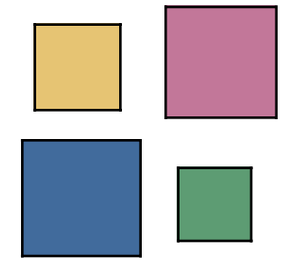
Martin Theus Department of Computational Statistics and Data Analysis, Augsburg University, Germany



Data Analysis Principles in Interactive Statistical Graphics

Principles of Data Analysis

- Quotes from John W. Tukey
  - on **EDA** :
    "… must be considered as an open-ended, highly interactive, iterative process, whose actual steps are segments of a stubbily branching, tree-like pattern of possible actions."
  - on **formalization**:
    "… the technology of data analysis is still unsystematized …"

    but also warns

    "Data analysis can gain much from formal statistics, but only if the connection is kept adequately loose."
- Little formalization has been done yet, …
- … still, certain principles are used over and over again.

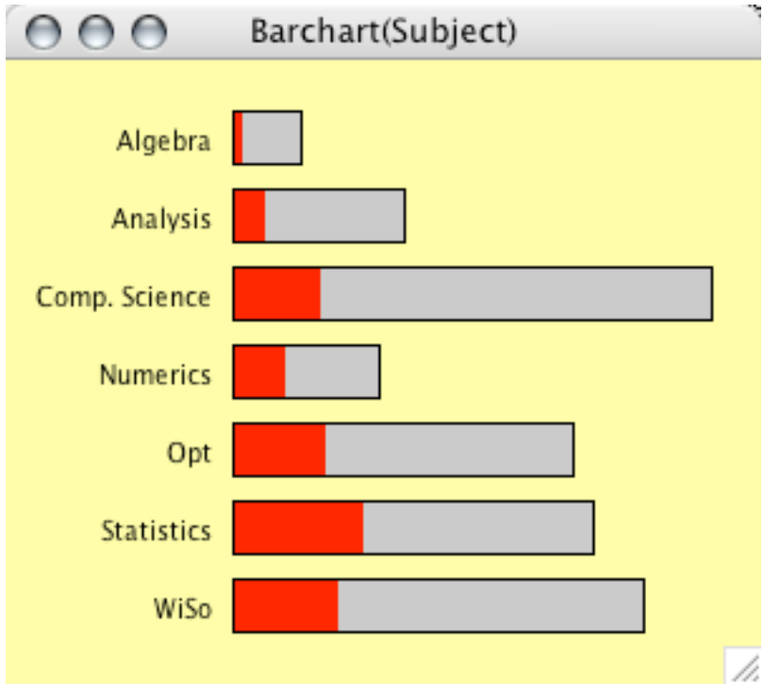
Plots in Mondrian: Summary

- Univariate Plots
  - Barchart / Spineplot
  - Histogram / Spinogram
  - weighted versions of Barchart and Histogram
- Bivariate Plots
  - Scatterplot
- Multivariate Plots
  - Mosaic Plot
  - Parallel Coordinate Plot
- Special Plots
  - Map
  - (SPLOM)

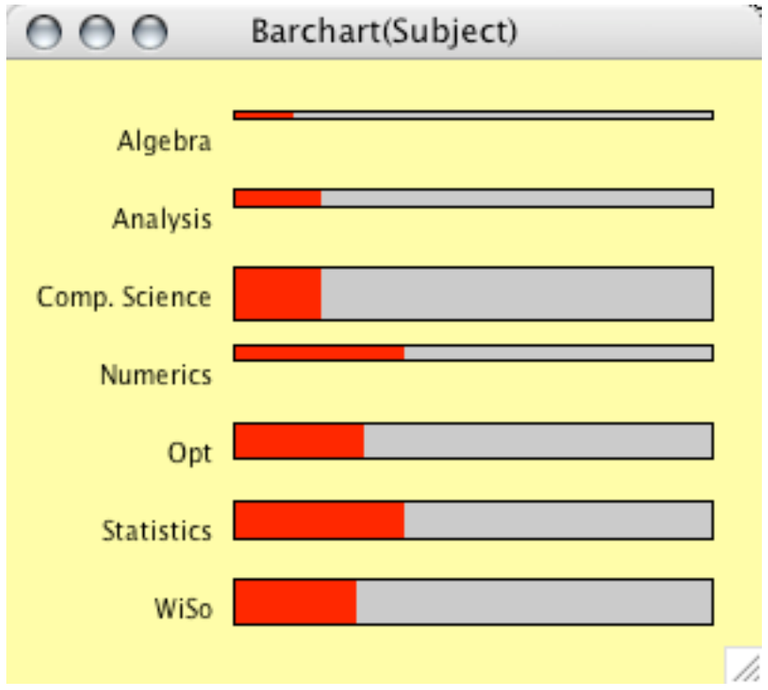
Barcharts / Spineplots

- Horizontal layout
- Sorting by
  - frequency
  - abs. / rel. highlighting
  - “hand”

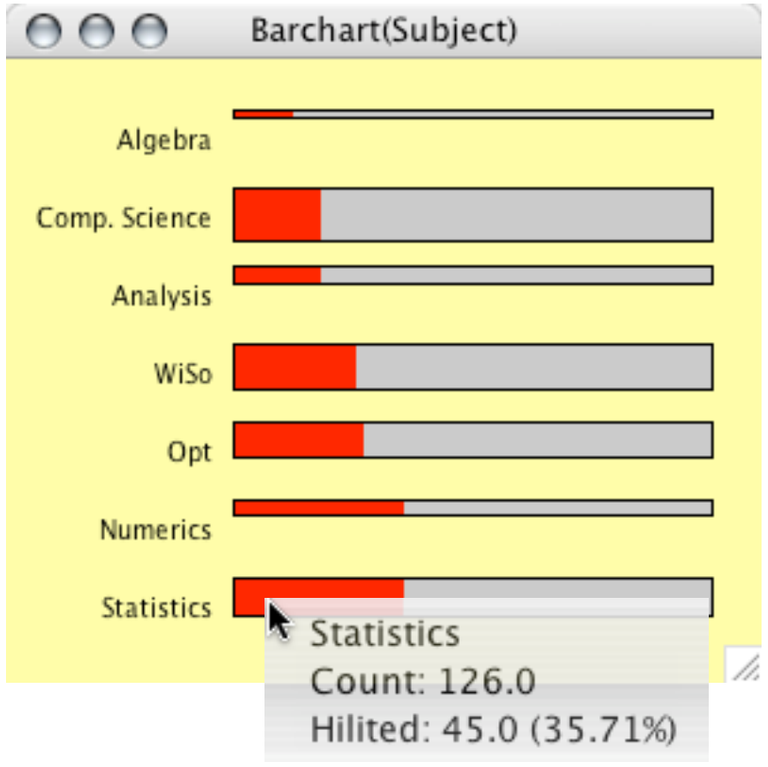
Barchart



Spineplot

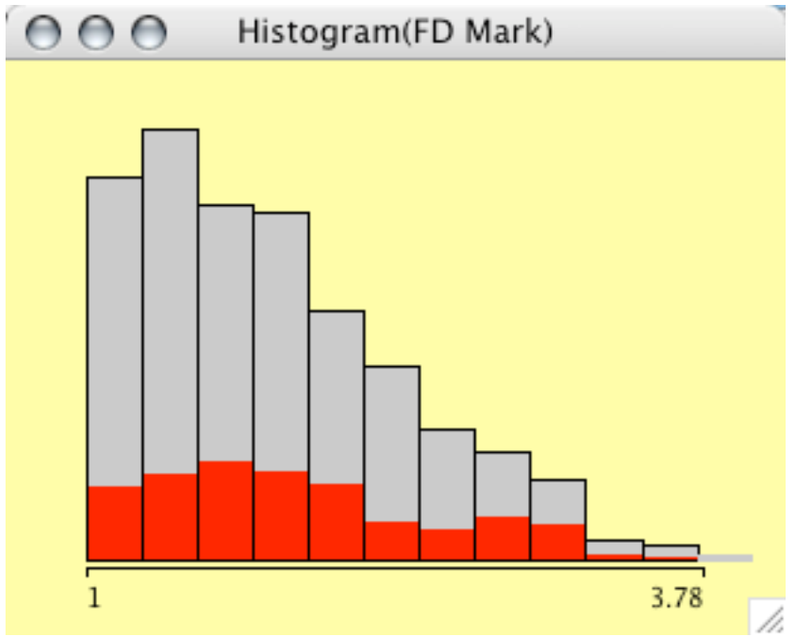


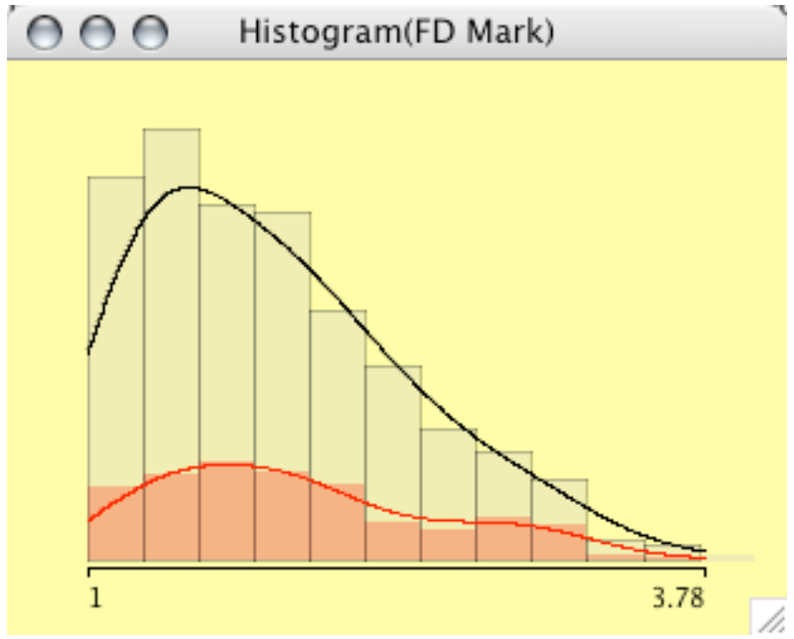
sorted Spineplot

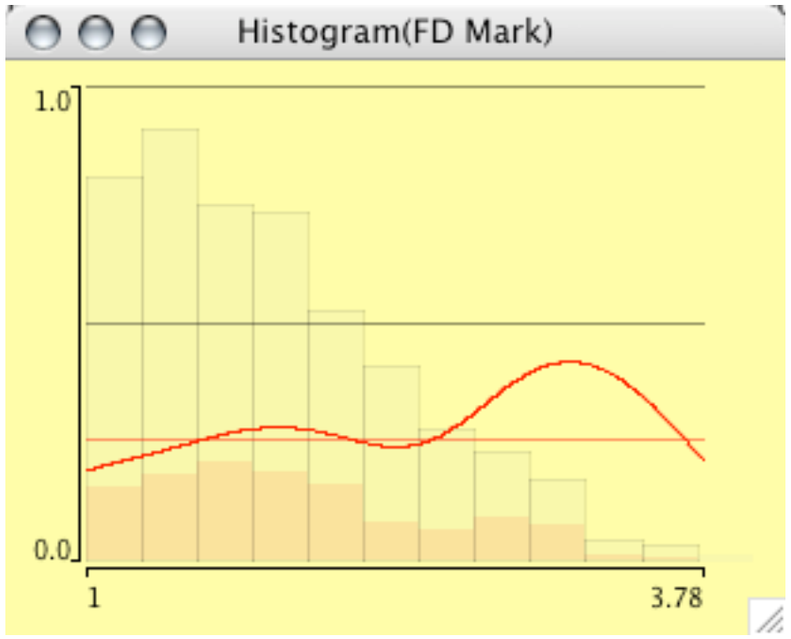


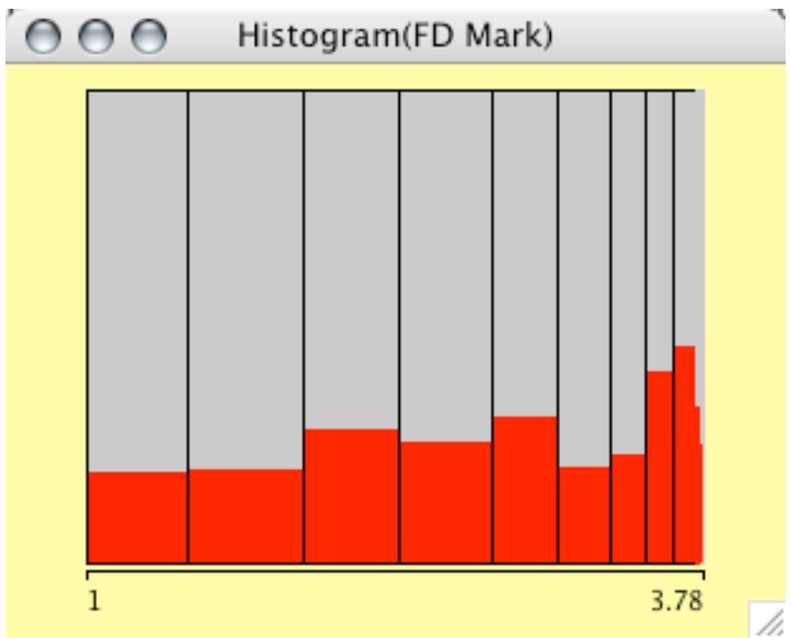
Histograms / Spinograms / CD-Plots

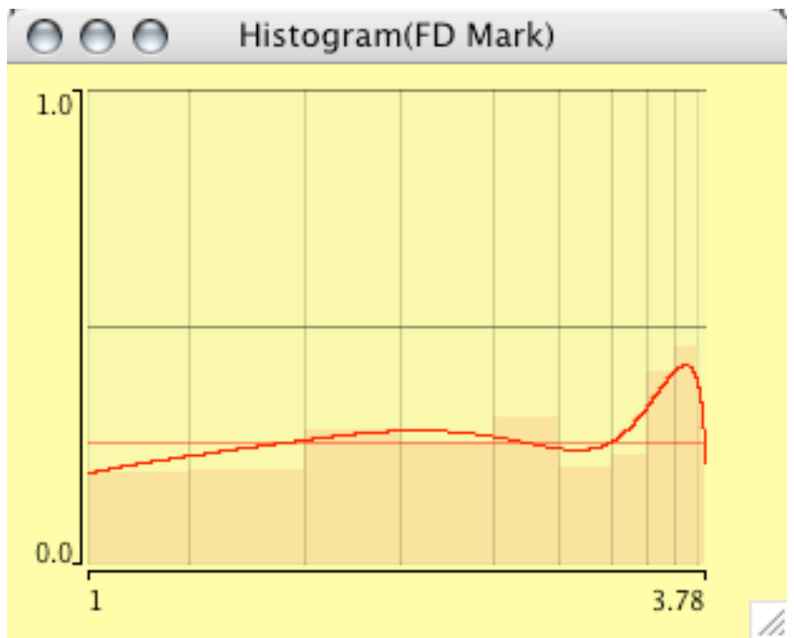
- Fully interactive (origin, bin width)
- Spinogram / Density / CD-Plot option







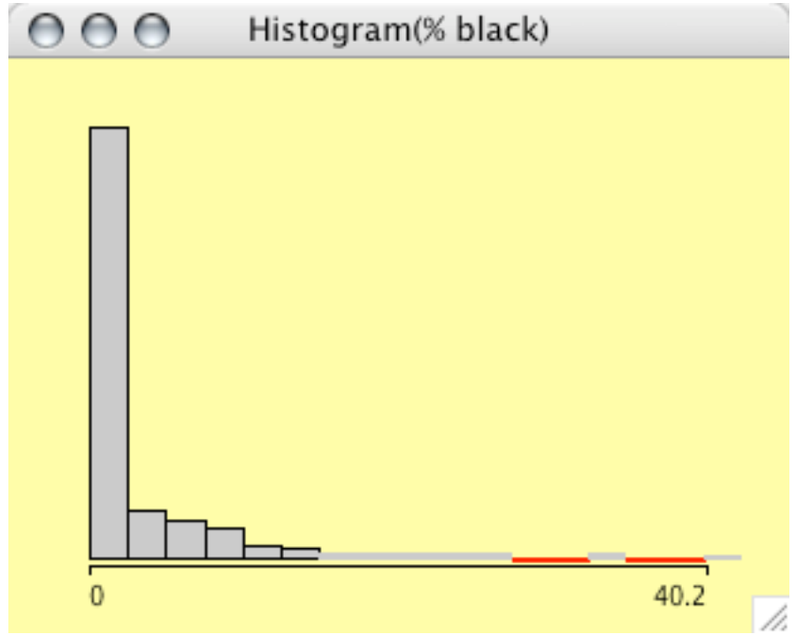




Weighted Plots

- Weighting plots is implemented for
  - Barcharts
  - Histograms
  - Mosaic Plots

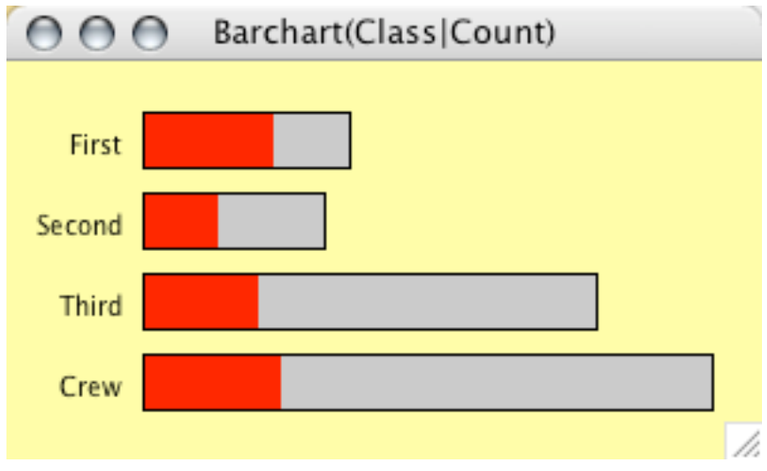
- Two versions of weighting
  - simple counts as weights for categories
  - two continuous variables (one >0!)

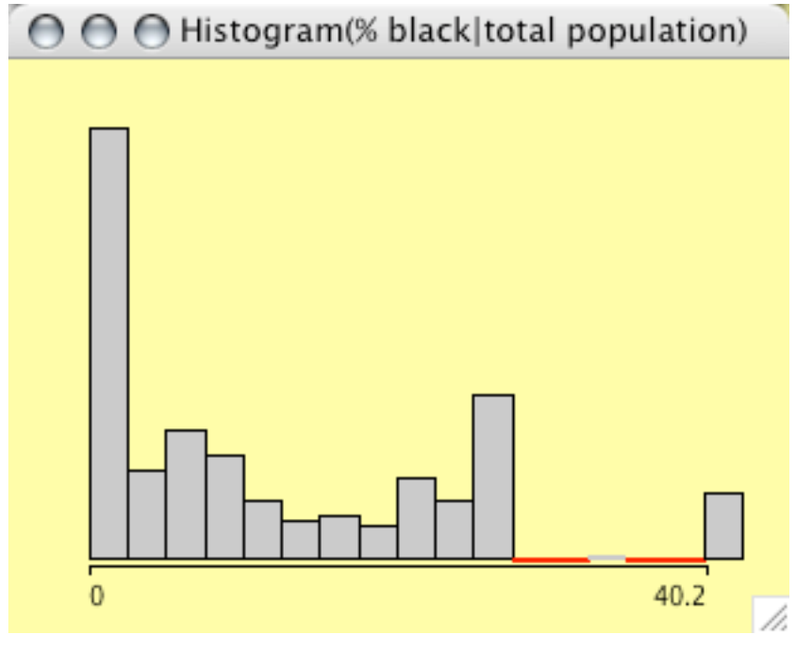


unweighted: counties

weighted: people

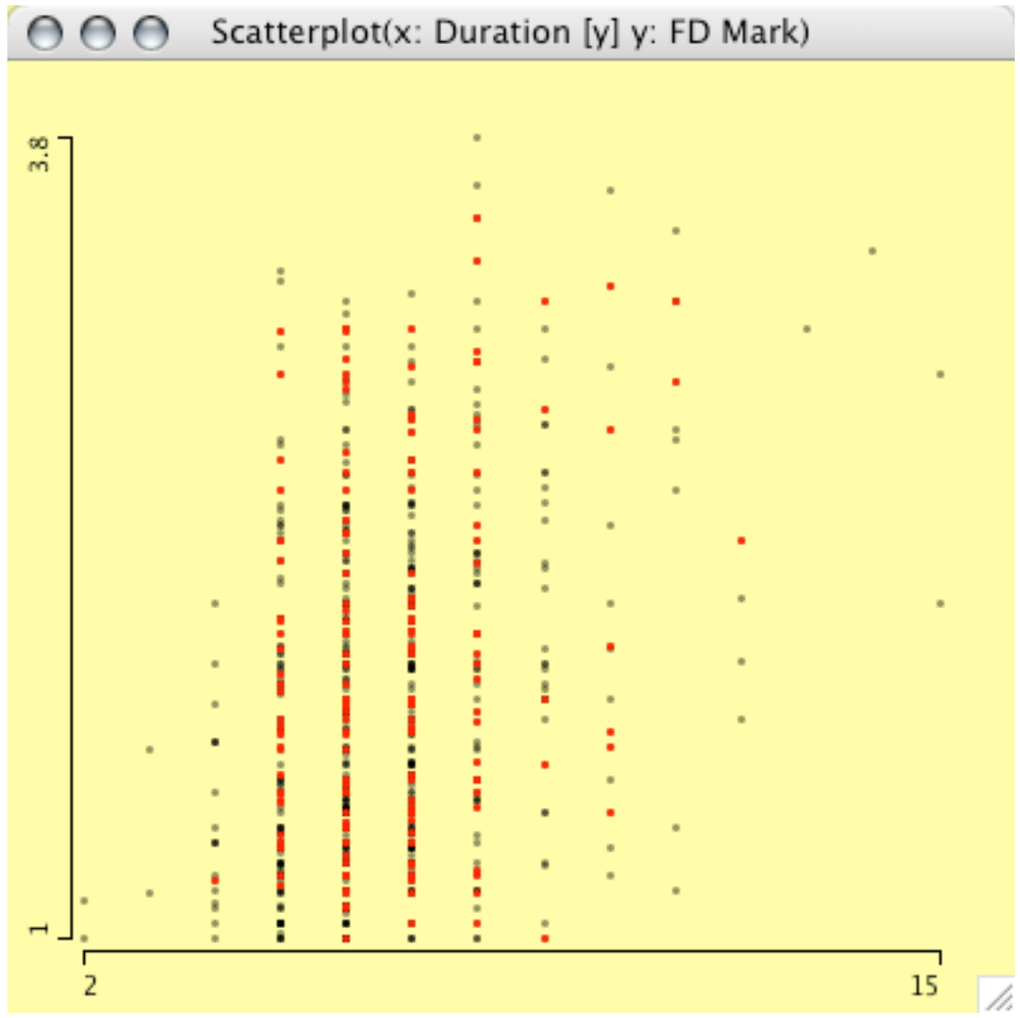
Weighted Categories

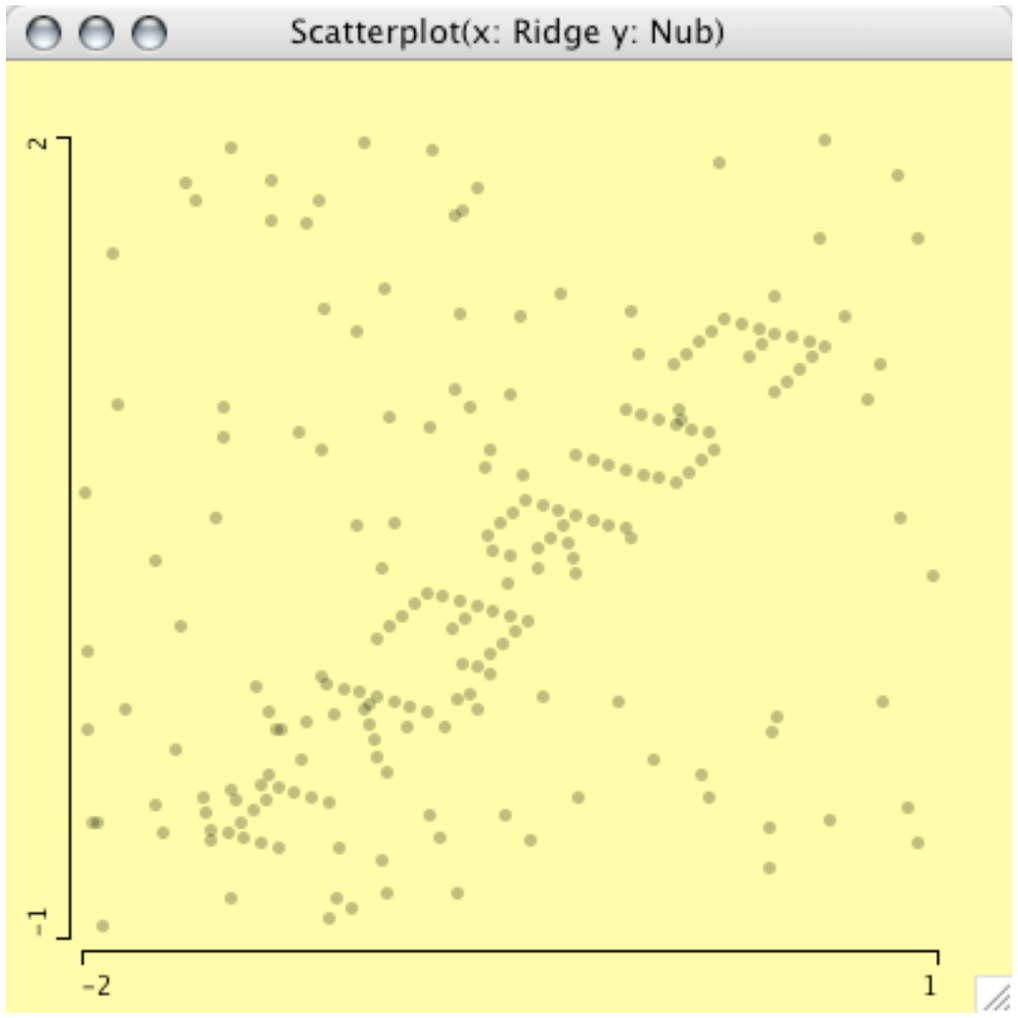




Scatterplots

- Supports
  - 3 levels of queries
  - zooming (hierarchical)
  - α-blending

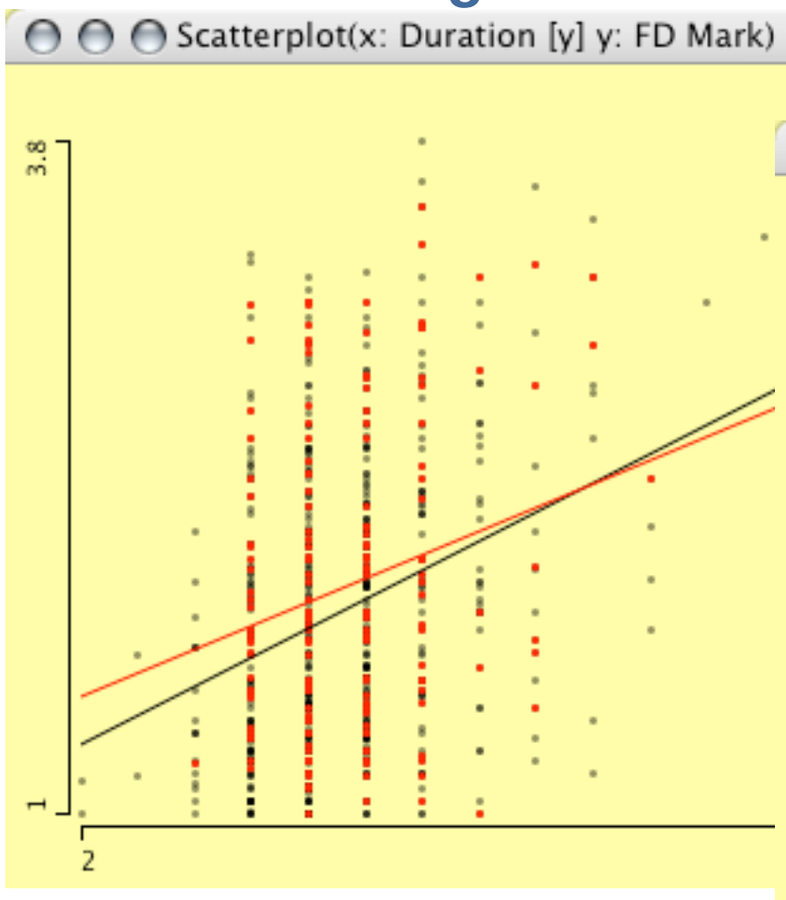




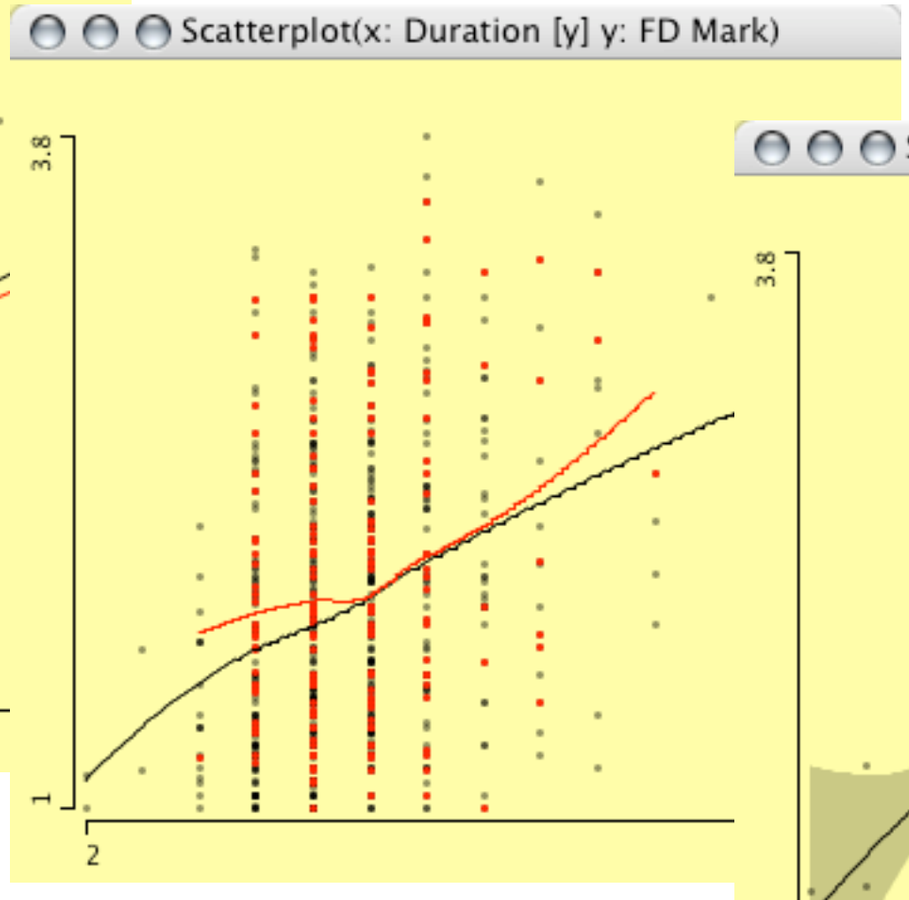
Scatterplot Smoother

- Scatterplot smoother enhance scatterplots to offer a better judgement of the mean function.

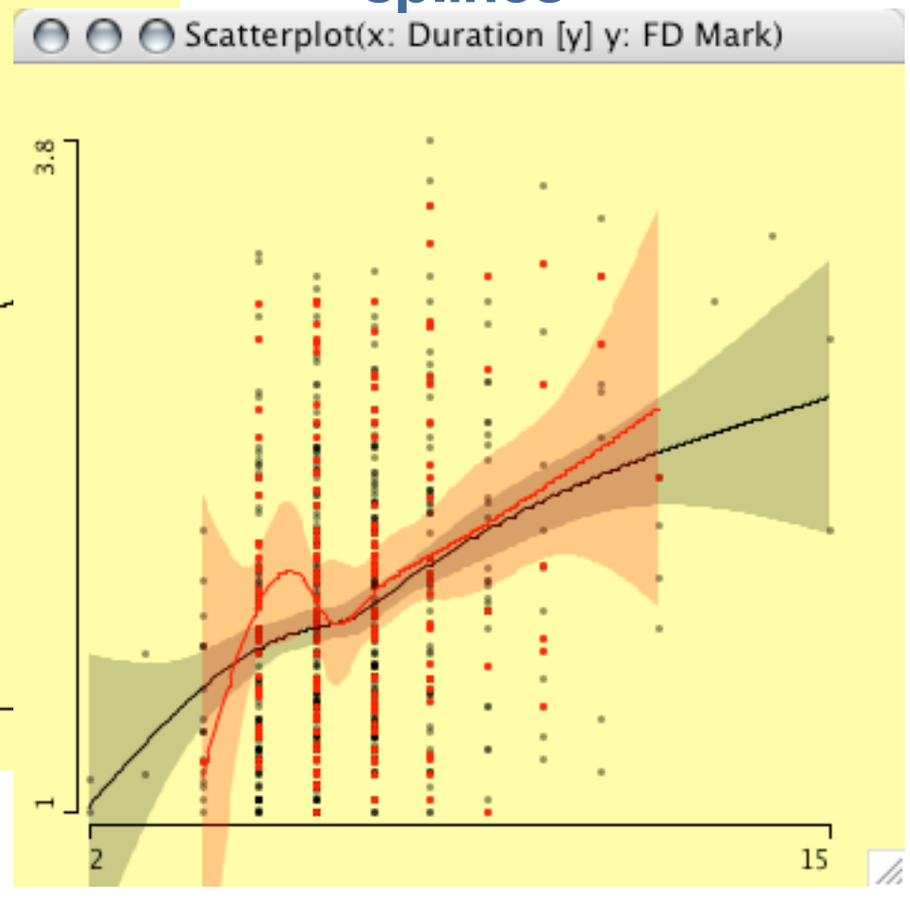
linear regression



loess

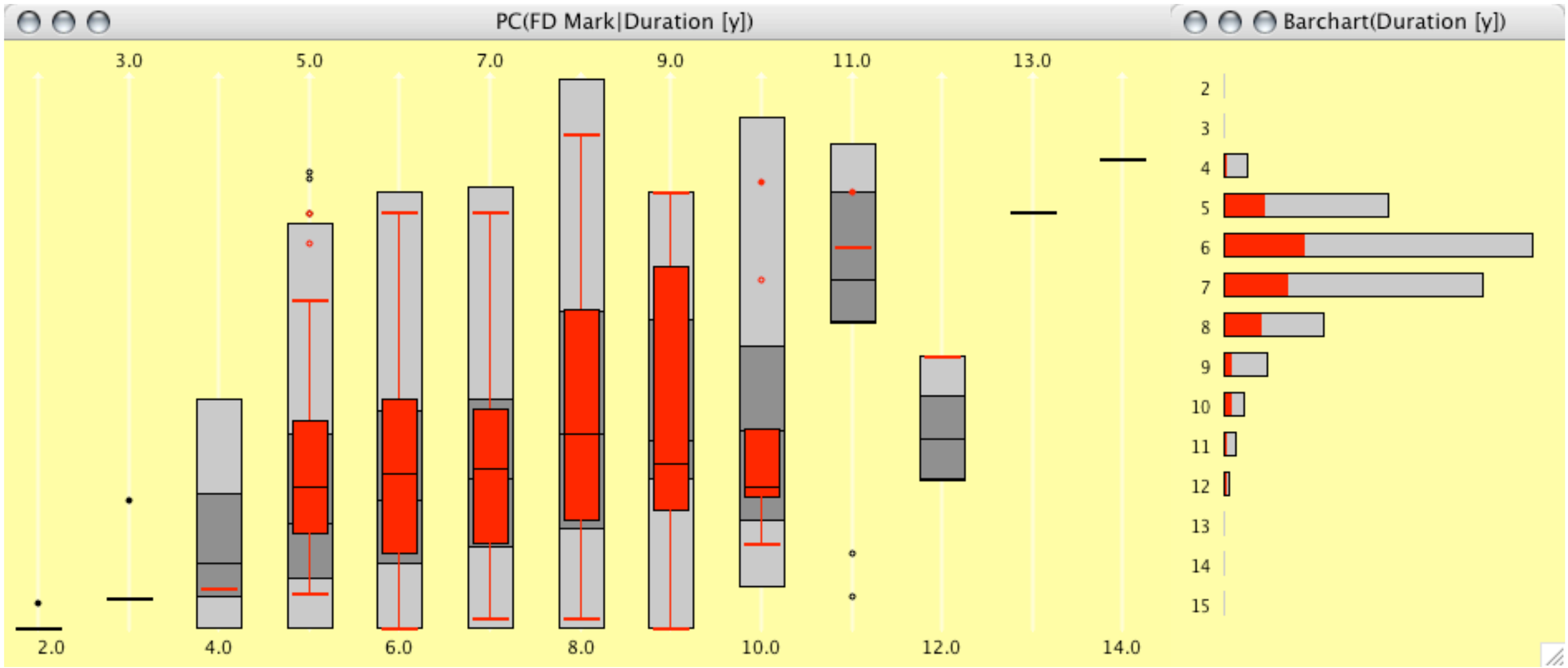


splines



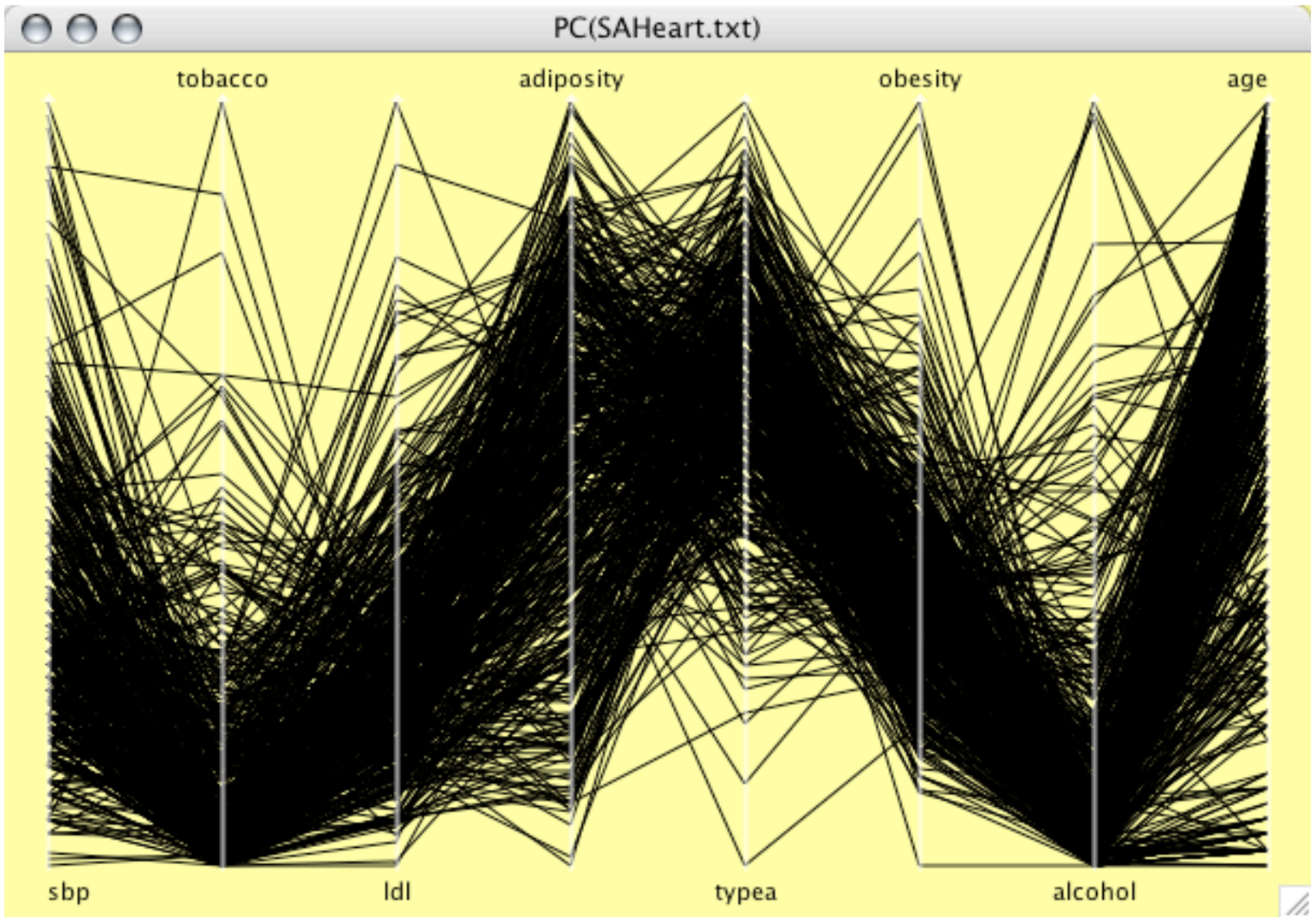
Boxplots y by x

- Most effective tool to explore first and second order moments
- Often confused with parallel boxplots and parallel coordinates



Parallel Coordinates

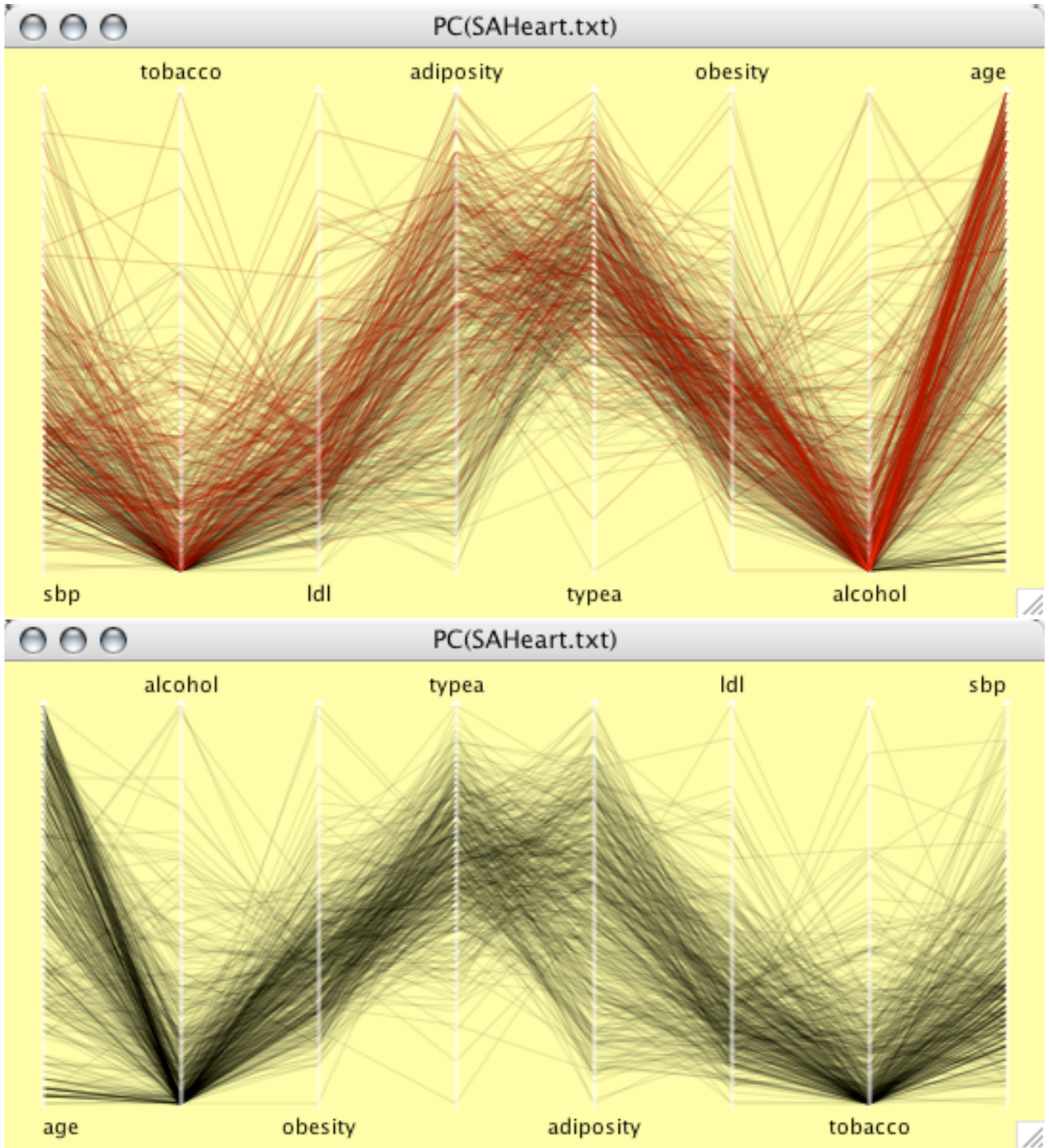
- Ideal tool to get a comprehensive view of a dataset
- Interactive options are crucial to make the plot usable
- Overplotting is even more serious than in scatterplots



Martin Theus Department of Computational Statistics and Data Analysis, Augsburg University, Germany

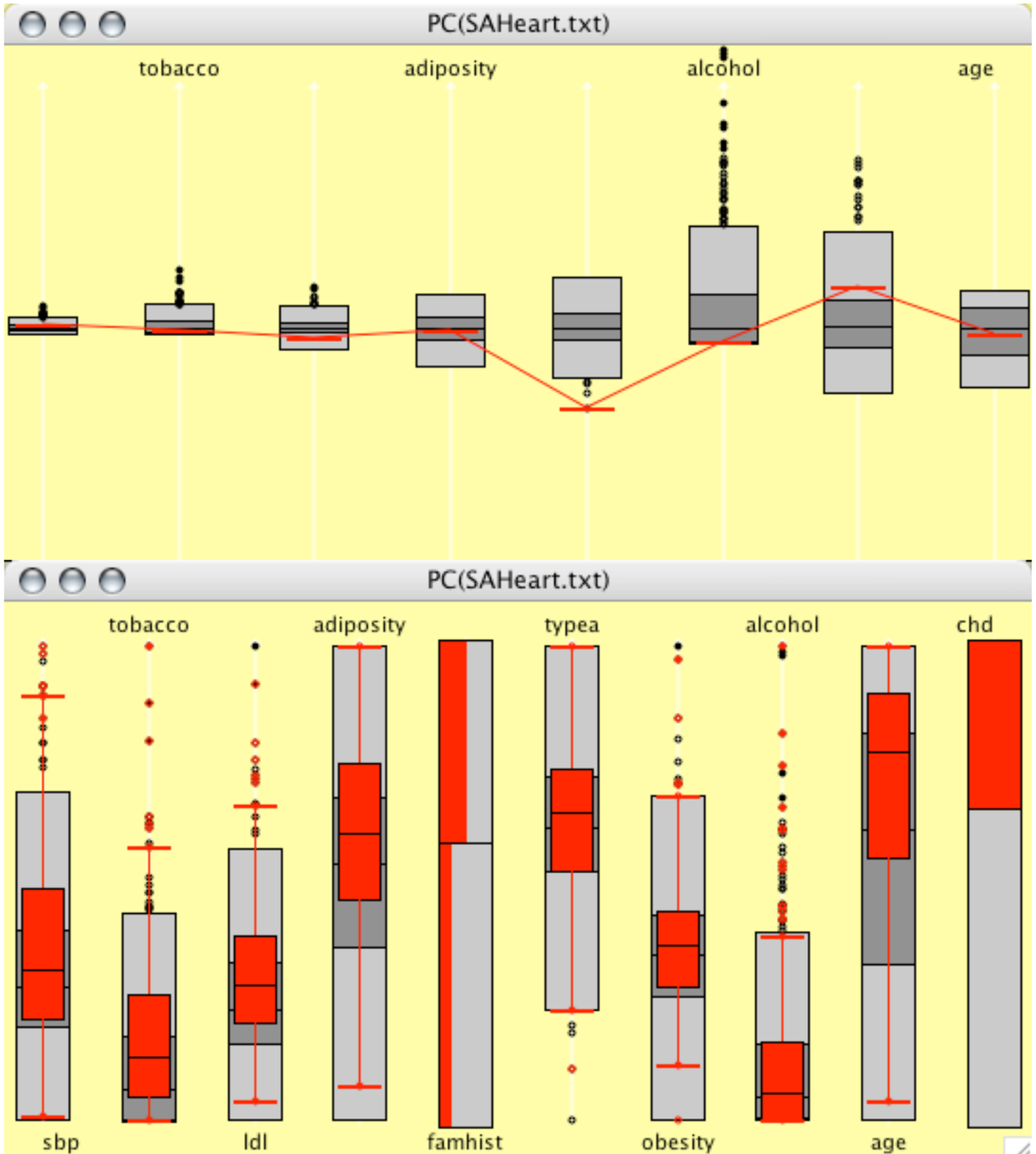
Parallel Coordinates: Interactions

- α-blending on data and selection
- Sorting of axes:
  - min
  - sdev
  - IQ-range
  - mean
  - median
  - max
  - *manual*
- Permuting axes to see **all** adjacencies of the axes.



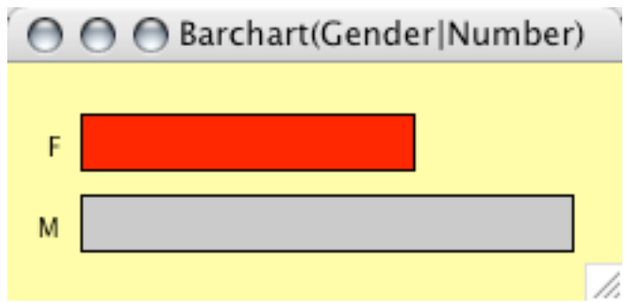
Parallel Coordinates: Variations

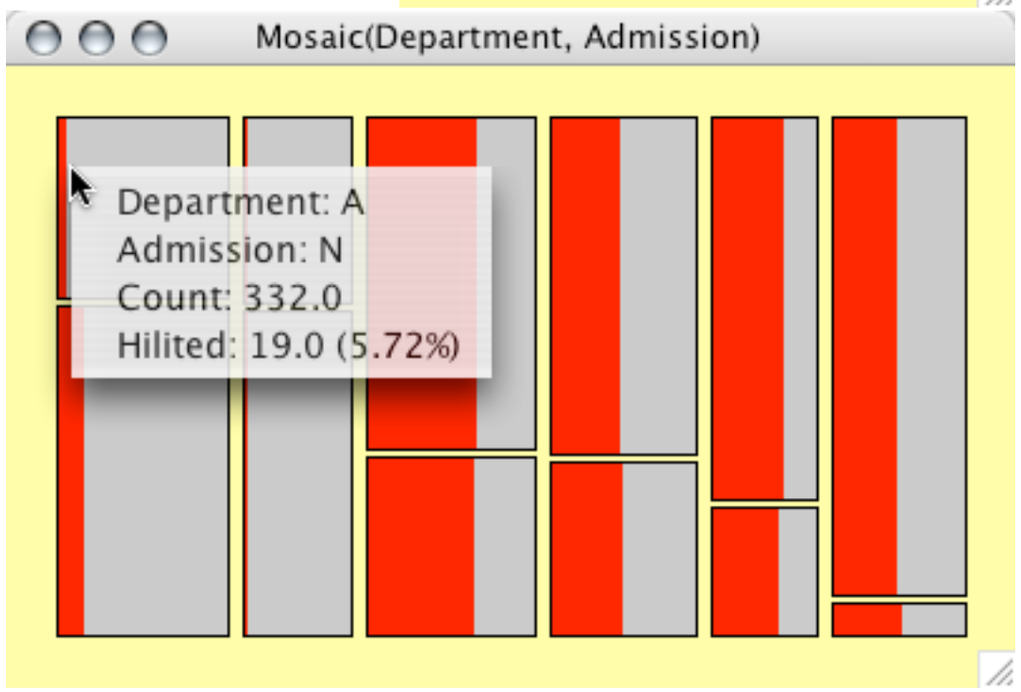
- Scaling options
  - individual
  - common
  - align at
    - mean
    - median
    - “value”
- Display options
  - Lines
  - Boxplots (includes categorical variables as spineplot)
  - Both (lines and boxes)
- Hot Selection
  - selected only
  - zoom to selection

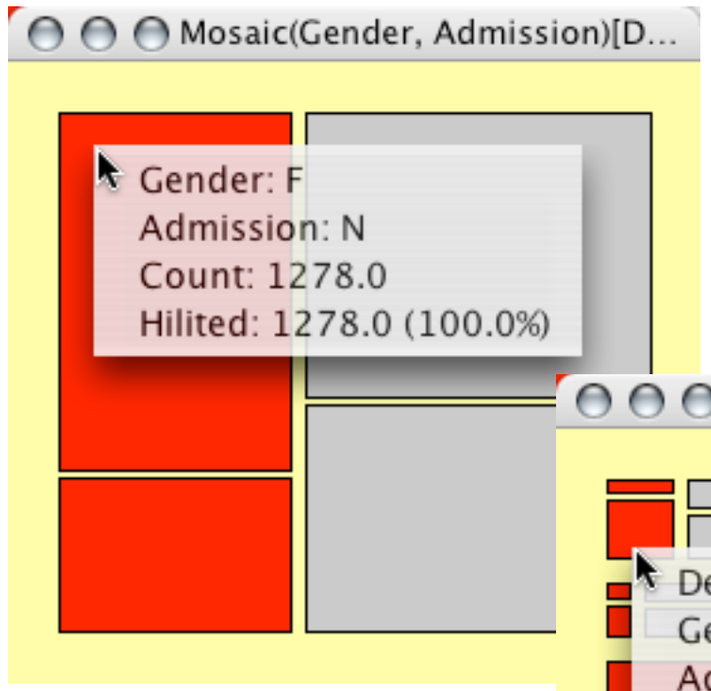


Mosaic Plots

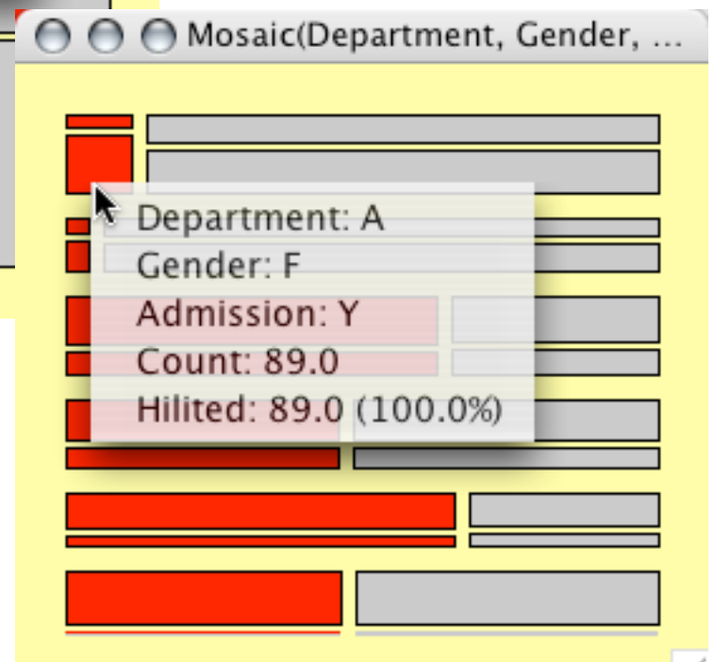
- Mosaic Plots generalize barcharts and spineplots in a **recursive** and **conditional** way to visualize high-dim. categorical data.
- Implementation is fully interactive (reordering, rotation, optional views, loglin-modeling, queries)
- Example:







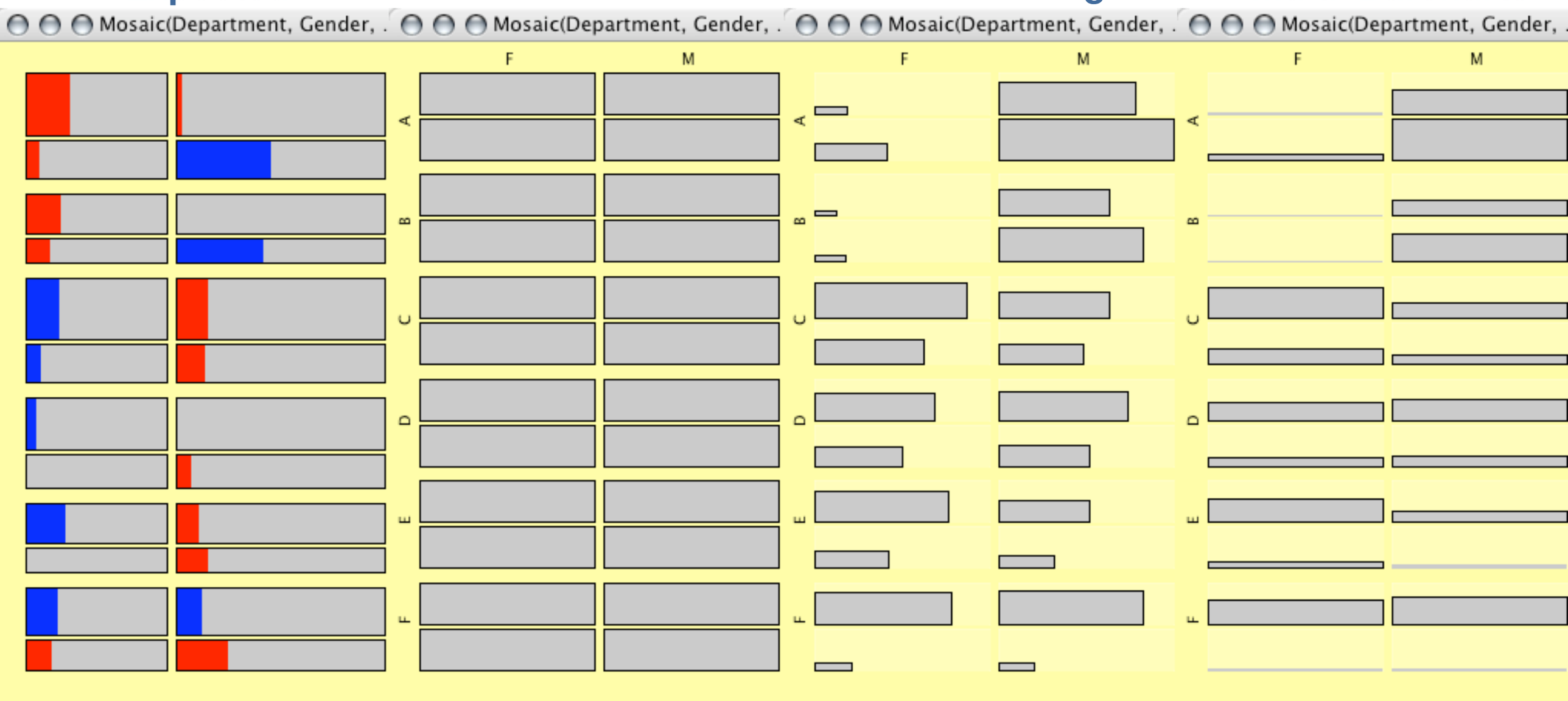
Simpson's Paradox



Mosaic Plots: Variations

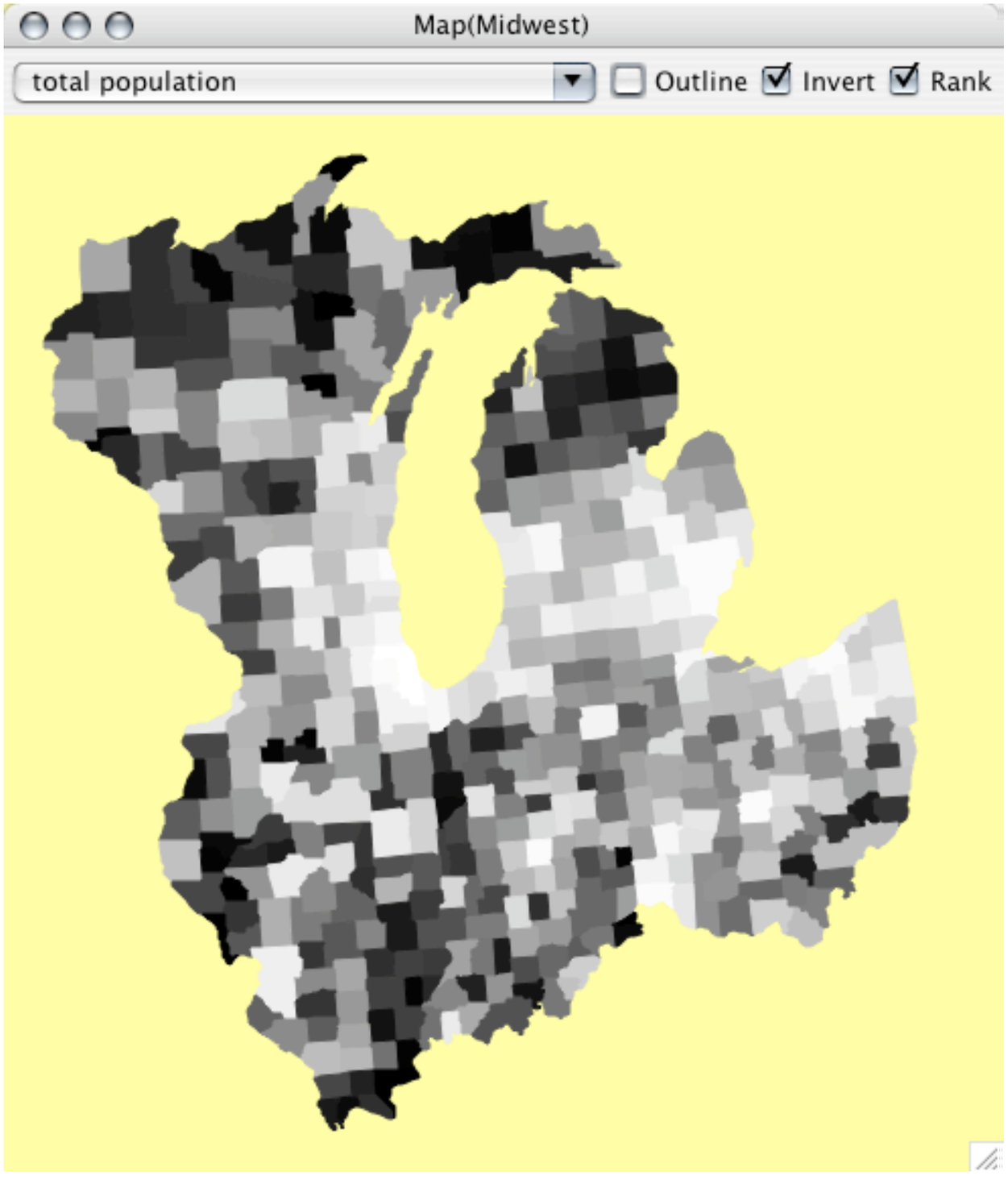
- Several variations of mosaic plots can be derived – some are very useful in 2-d, some need the "right" aspect ratio

Expected **Same Bin Size** **Fluctuation Diagram** **Multi. Barcharts**



Maps

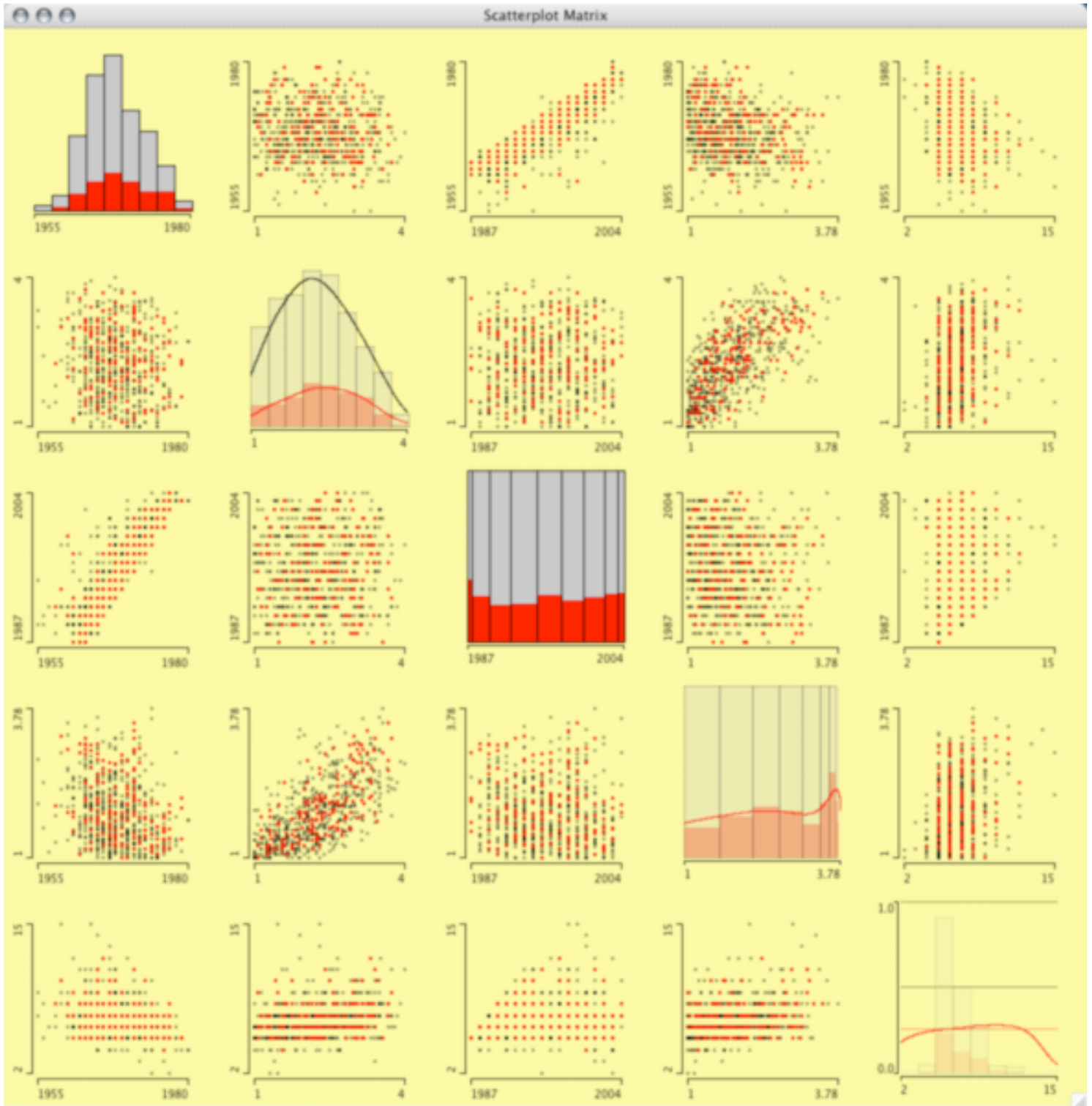
- Maps are fully integrated into the linked highlighting mechanism
- Zoom is available
- Extended queries
- Options are
  - outline
  - linear choropleth maps
  - quantile choropleth maps
  - change sign of color scale

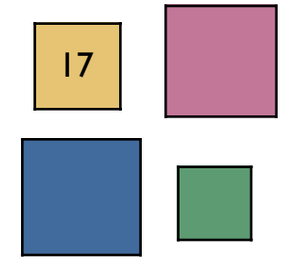


Scatterplot Matrix – SPLOM

- Still experimental …

  … but a very nice addition to parallel coordinate plots.





Some Strategies

- Strategies in interactive statistical graphics can most easily be derived from classical statistical pendants.
- Some are implemented in plot variations itself, cf. spineplot etc.
- Others can be found in specific plot ensembles, which can visualize certain patterns.
- To gain more statistical support, graphics can be enhanced with statistical estimates and tests
- Strategies should not become too restrictive
  ⇒ Exploration is what makes interactive graphics so powerful!

2 dim. Interactions

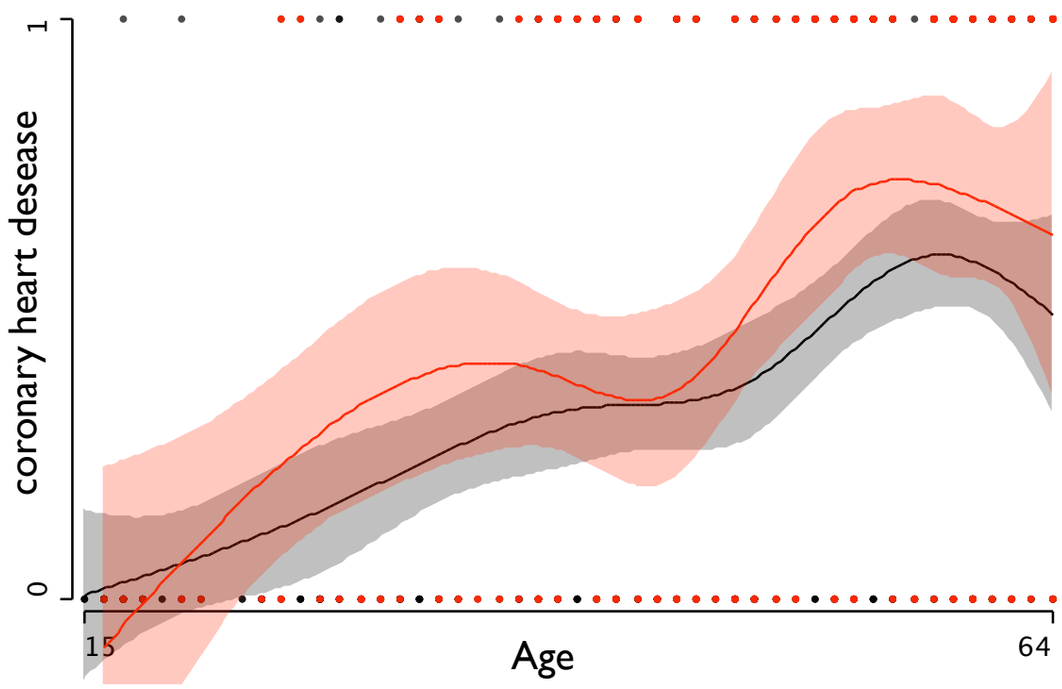
- 2 categorical variables ⇒ Mosaic Plots ✔

  (associations can be visualized via plotting expected values of models in question)

- 2 continuous variables ⇒ Scatterplots ✔

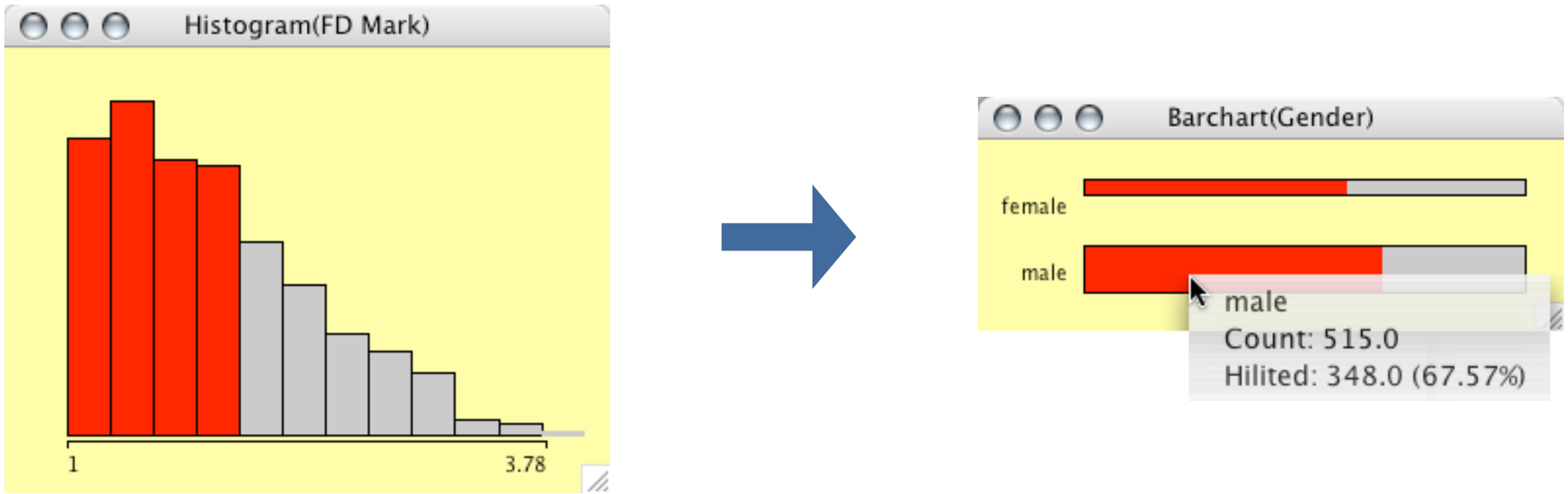
  (if the role of independent and dependent variable is known, scatterplot smoother can enhance the graphics)

- Sometimes plotting categorical variables in a scatterplot makes sense



2 dim. Interactions

- Mixed Scales
  - continuous ⇒ categorical



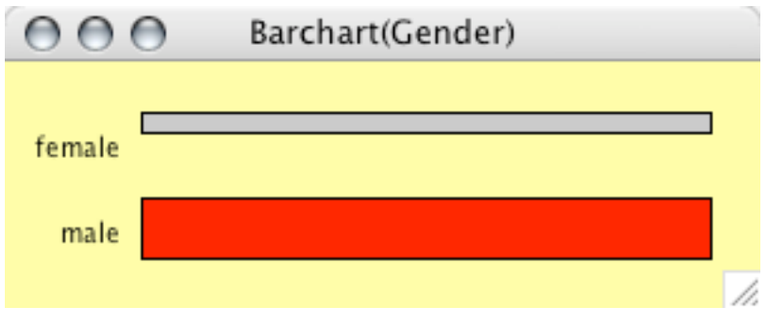
read:
“given gender is ‘male’, what is the probability of having a mark better than 2”

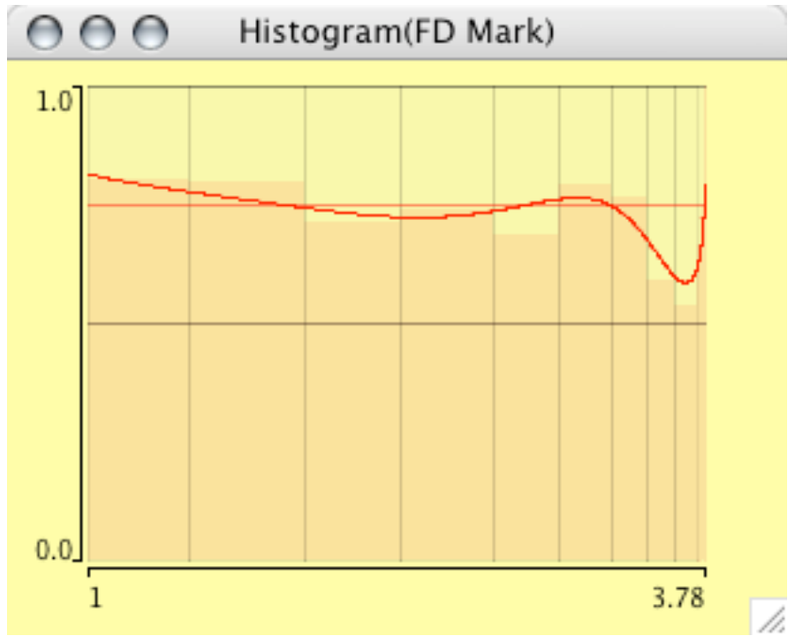
or

$$\mathrm{P}(mark < 2 \mid gender = \text{‘male’})$$

2 dim. Interactions

- Mixed Scales
  - categorical ⇒ continuous





read:
“given a mark better than 1.25, what is the probability that gender is ‘male’”

or

$$\mathrm{P}(gender = \text{`}male\text{'} \mid mark < 1.25)$$

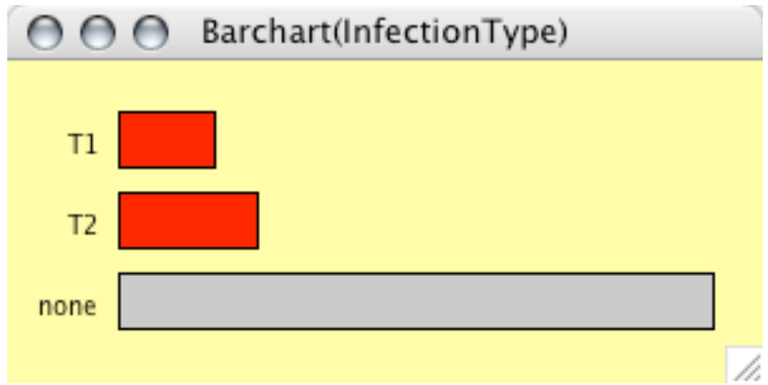
or, get the global estimate via density estimators.

Categorical Response Models

- Situation:
  Multivariate categorical inputs and univariate categorical output

Inputs

Output



Given the mother was in a certain risk group, what is the infection rate?

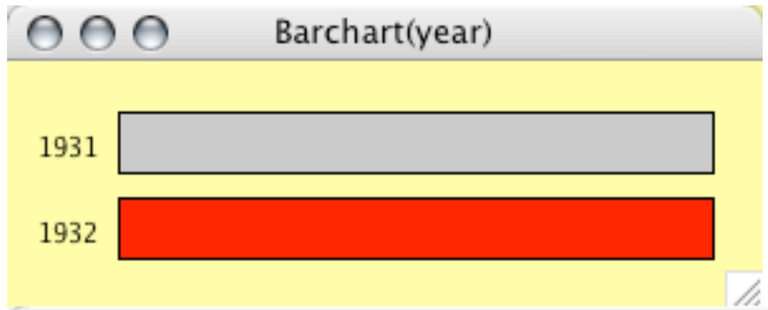
ANOVA

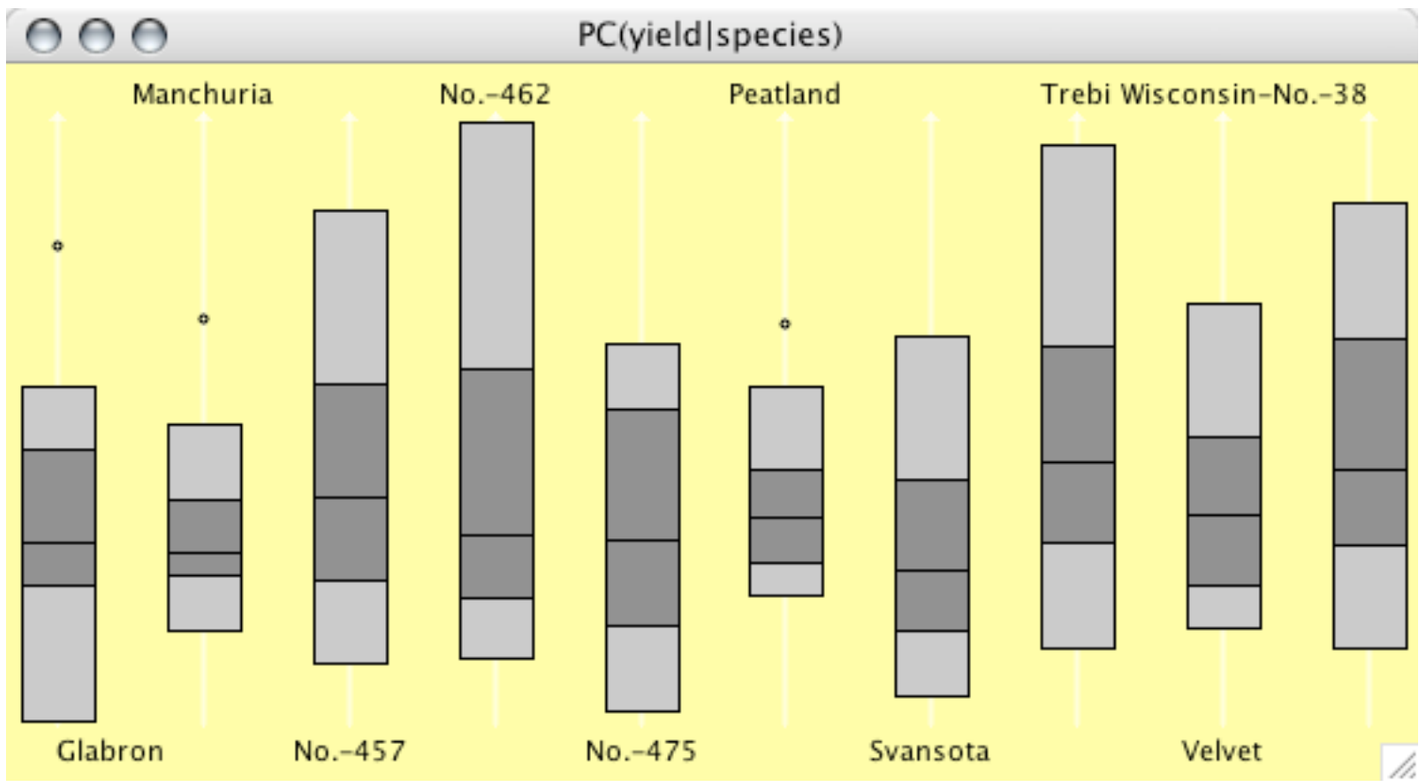
- Example: Barley Data
  - One way

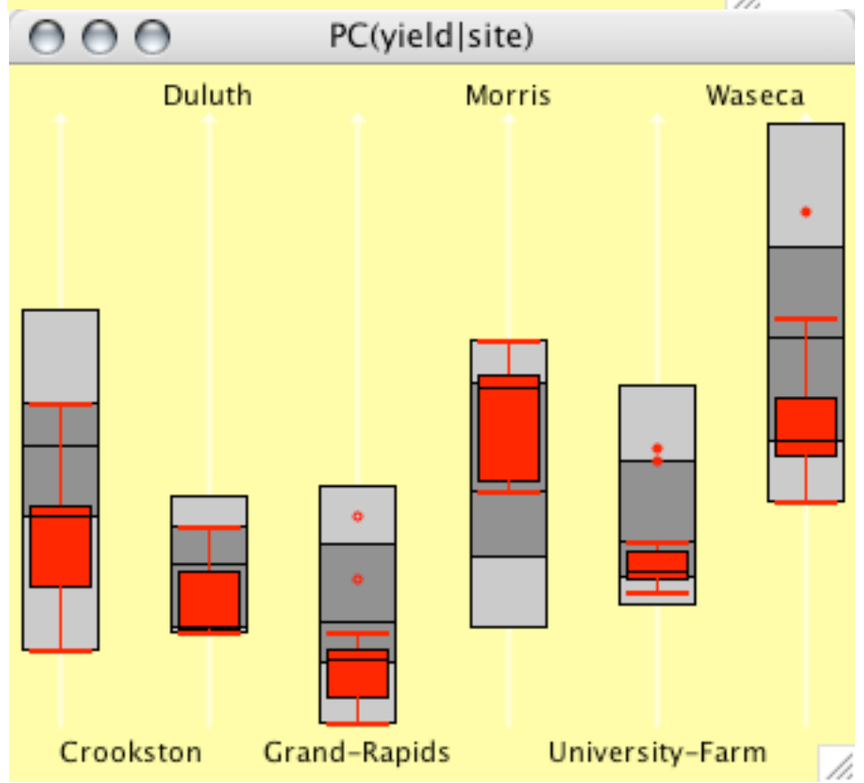
    `yield ~ species`

  - Two way

    `yield ~ year * site`







Let's "play" with the data ...

- Math Students Data:
  Data from Math students at Augsburg University over the last 18{} years
  - Gender male, female
  - Year of Birth
  - School Place s=Swabia, b=Bavaria\s, d=Germany\{sUb}, r=World\{sUbUd}
  - Maj. Subject Business Math or pure Math
  - Start year of start
  - 1stD Month month of "first degree"
  - 1stD Year
  - 1stD Mark
  - Thesis Mark Mark on the final thesis (diploma)
  - FD Month month when final degree was finished
  - FD Year
  - FD Mark
  - Summer Start did student start in Summer?
  - Duration [y]
  - Subject area of specialization for thesis

Conclusion

- A variety of plots is needed for a comprehensive exploration of a data set.
- The key feature is the comparison of subgroups via linked highlighting.
- Interactivity offers many views and “what if” scenarios.
- Adding R-functions can enhance the graphics.
- Some standard statistical procedures translate into graphical analyses.

- **BUT**, exploratory and graphical methods will remain a niche market, as long as they are not picked up in teaching

  ☝ the tools are here!