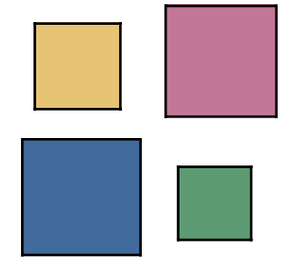
Workshop on Modern Statistical Visualization and related Topics (5)
Recent developments in Mondrian and R



3 Talks on Interactive Statistical Graphics

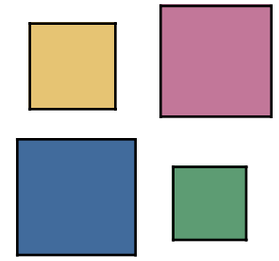
I. Design Principles

II. Application in Data Analysis

III. Implementation Issues

(All talks use Mondrian for illustration)

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



Designing Interactive Statistical Graphics Software

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

Exploratory Data Analysis (John W. Tukey)

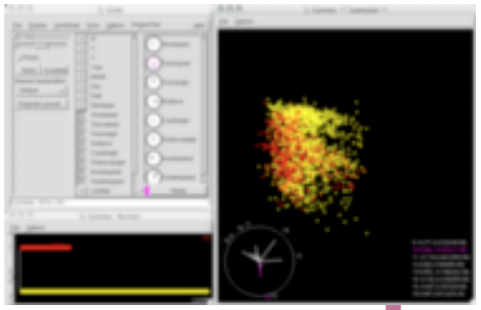
- **Visual Exploration**
  Usually the goal is to identify qualitative information rather than the quantification of an “effect”
- **Extensive User Interaction**
  In most cases the in no single best method or tool to explore the data. Thus the user needs the flexibility to interact with the data, statistics and graphics
- **Re-expressing**
  Dynamic transformations like:
  - Box-Cox
  - Elimination of outliers

  must give many alternative views on the data, almost instantaneously
- **Support for Meta-Data**
  Often the data matrix itself is not sufficient to cover all necessary information to understand the data and problems

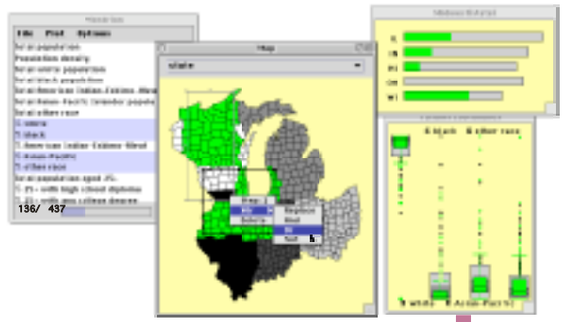
Interactive Tools: History

1983
SPLOM
Becker et al.

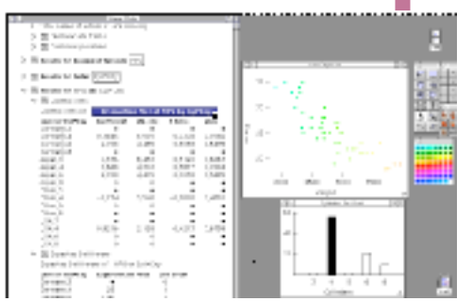
1999
ggobi
Swayne et al.



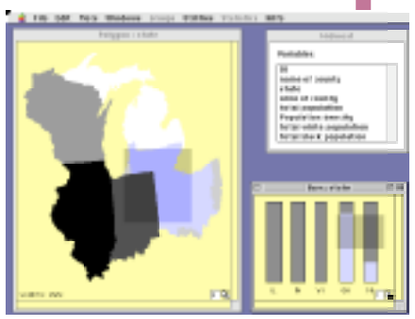
1997
Mondrian
Theus



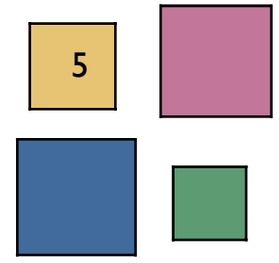
1973
PRIM-9
Tukey et al.



1985
DataDesk
Vellemann



1993
MANET
Unwin et al.



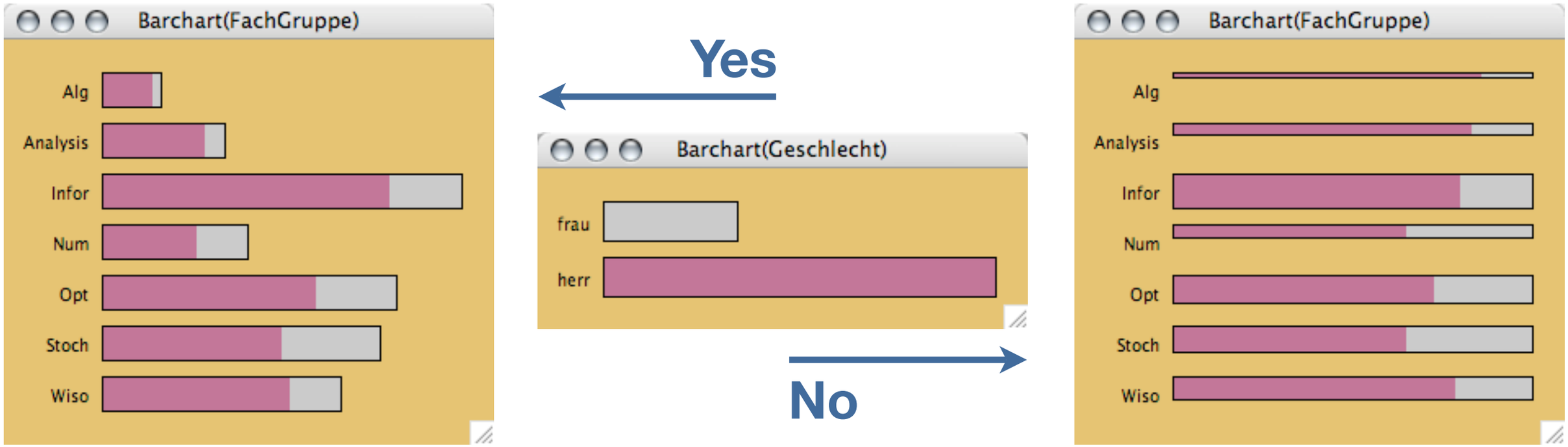
Elements of Interactive Statistical Graphics

- **Interaction**
  - Selection with linked highlighting
  - Modification of plot parameters
- **Interactive Graphics ≠ Dynamic Graphics**
- **The 4 pillars of ISG**
  - **Highlighting**
    highlight a selected subgroup
  - **Selection**
    selection of a subgroup of interest
  - **Query**
    query information on objects for non-obvious information
  - **Warnings**
    point the user to potentially misleading information

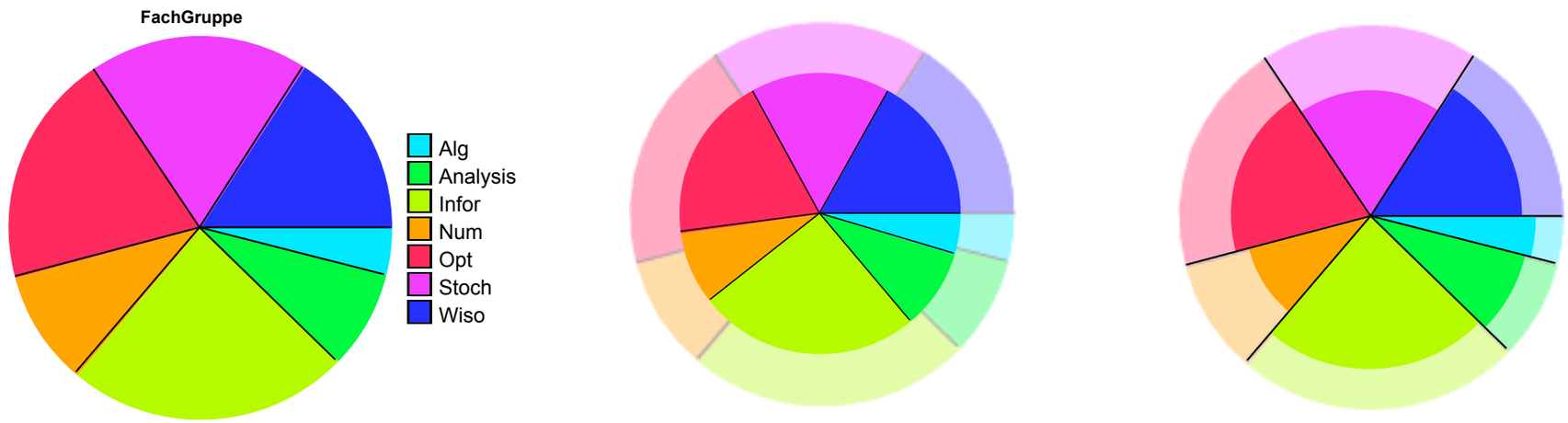
(– **Dynamic Variation**
  instantaneous change of plot parameters)

Highlighting

- **Question**: Is the highlighting of the same kind as the plot itself?



- But:



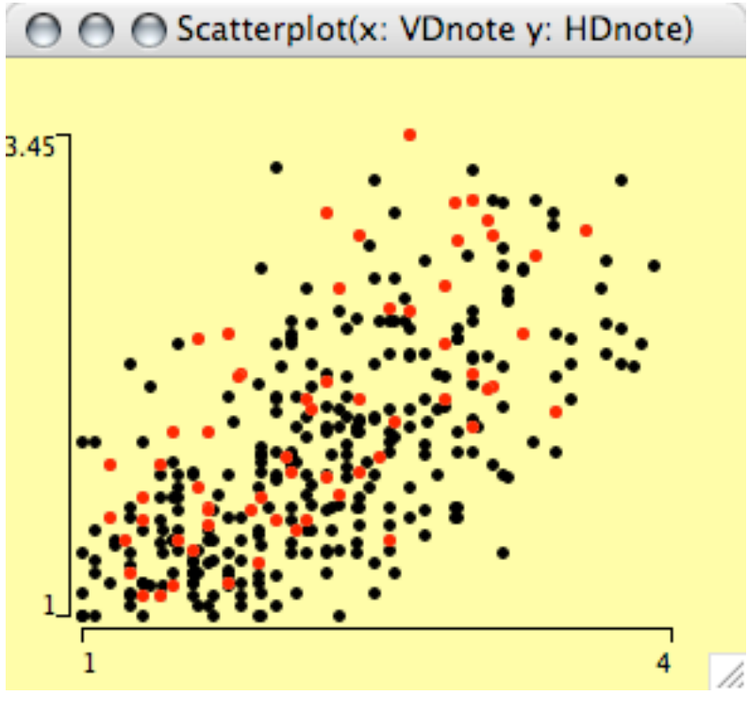
Highlighting: Complement vs. All

- **Question**: To what should the selected subset be compared?

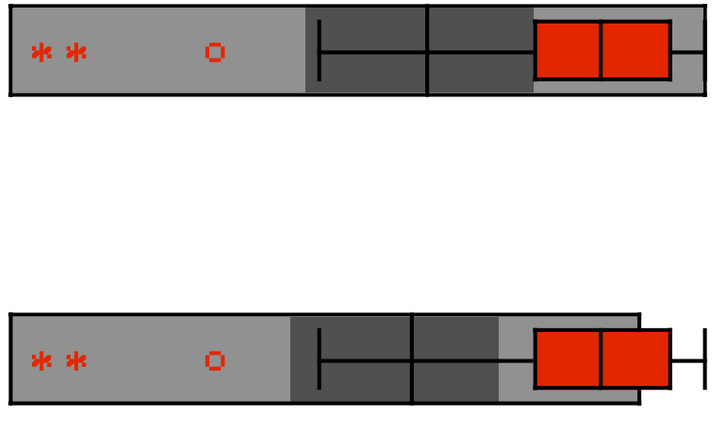
  (a) the complete sample or
  (b) the complement of the selection

  – For many plots it does not matter, e.g. scatter plots, parallel coordinate plots …
  – Some plots are different, e.g. box-plot

Scatterplot

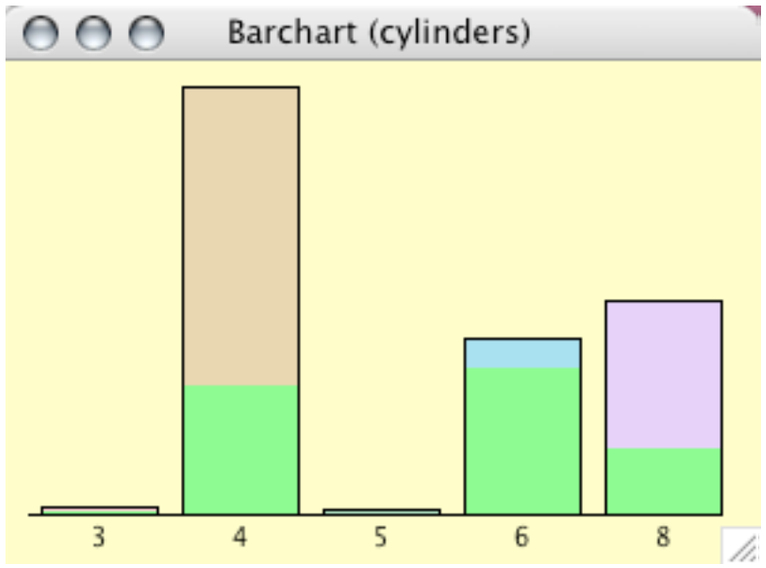


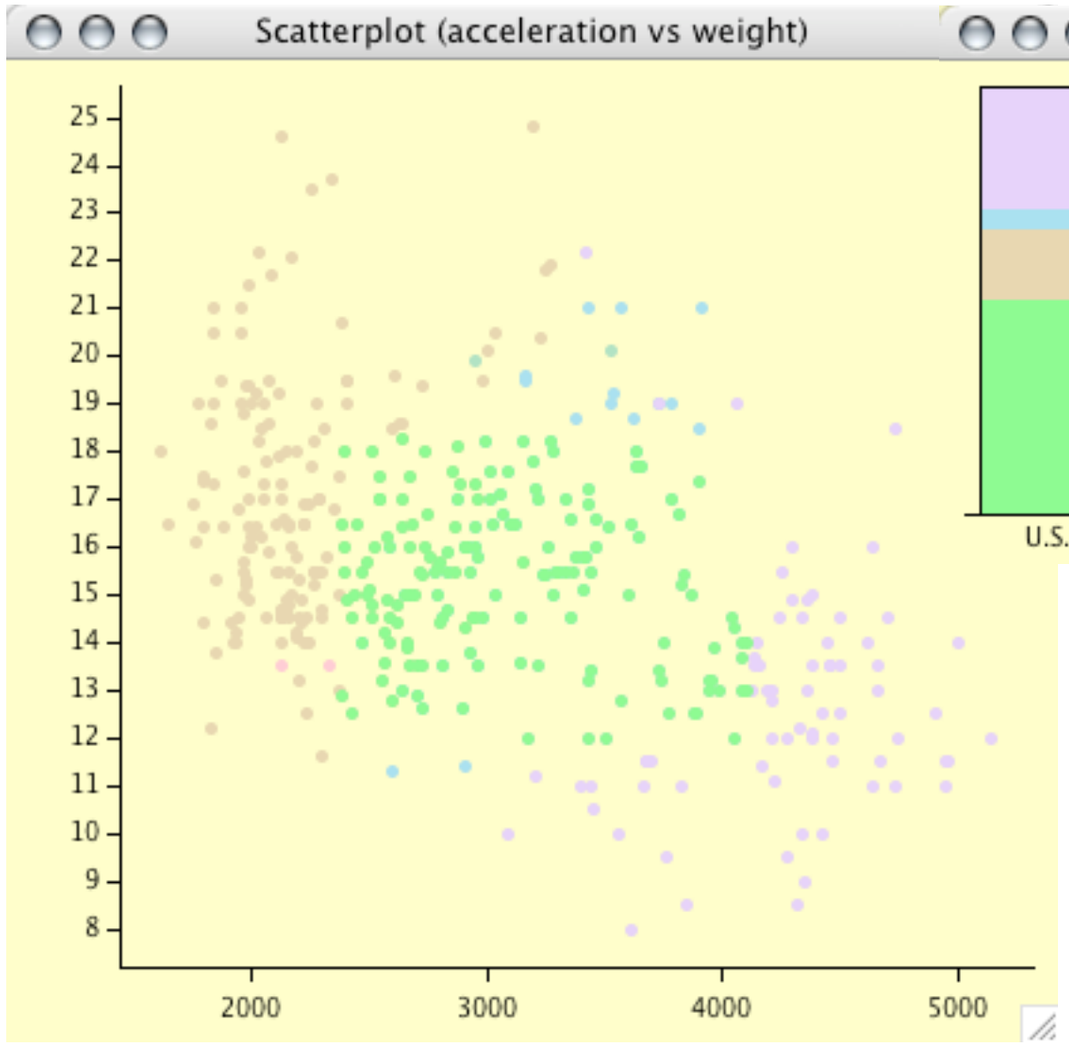
Box-Plot

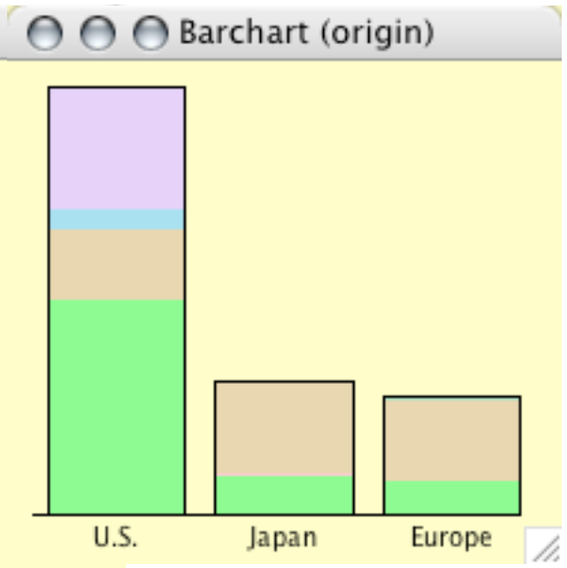


Highlighting vs. Color-Brushing

- Selection and Highlighting essentially divide the data into two groups, i.e. a virtual binary variable is introduced
- In settings where we want to compare more than two groups simultaneously color brushing can be used …



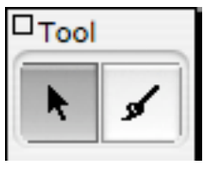




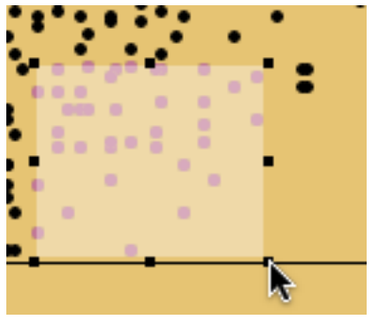
… with some limitations!

Selection

- Different Tools can be provided to select data:
  - **Pointer**
    The Pointer is used to select single points.
  - **Drag-Box**
    The Drag-Box selects retangular regions in a graphics window.
  - **Brush**
    Brushing allows a dynamic change (movement) of the selected region – usually a rectangle.
  - **Slicer**
    The slicer selects intervals along an axis dynamically.
  - **Lasso**
    The lasso allows the most flexible definition of the selection area. Startpoint and endpoint are always connected.



MANET

DataDesk

Mondrian

Selection: History

- **Standard**

  New selections replace the old selection set. (e.g. XGobi)

- **Advanced**

  Existing selections can be combined with further selections via AND, OR, XOR, NOT, … (e.g. DataDesk)

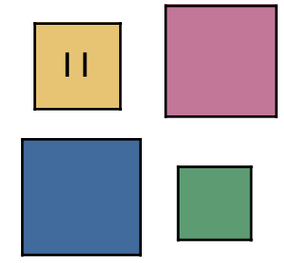
- **Selection Sequences**

  Any element in a hierarchical sequence of selections is stored and can be altered. (e.g. MANET)

- **Smart Selection Sequences**

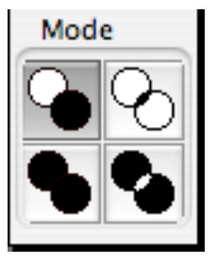
  Arbitrary number of selections per plot, which are invariant towards manipulations of the plot. (e.g. Mondrian)

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

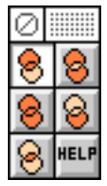


Selection: Modes

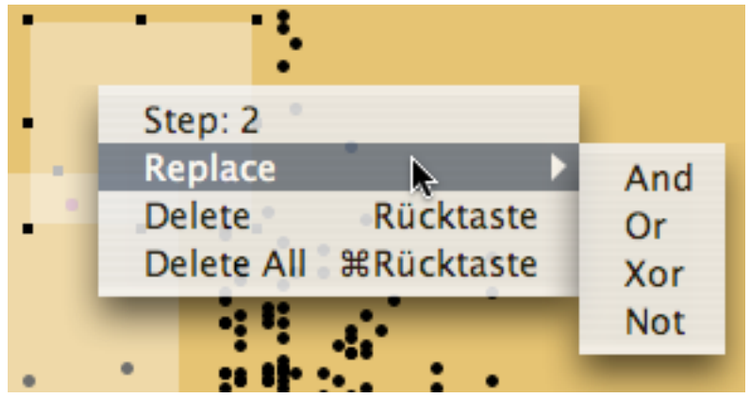
- Simple / Standard / Default
  - Only points in the selected region are selected.
- Intersection / AND / ∩
  - Only points that already were selected and are within the new selection stay selected.
- Union / OR / ∪
  - The newly selected points are added zu the current selection.
- Toggle / XOR / ⊕
  - Selected points are deselected, unselected are selected.
- Negation / NOT / ¬
  - Points in the selection region are taken out of the current selection set.



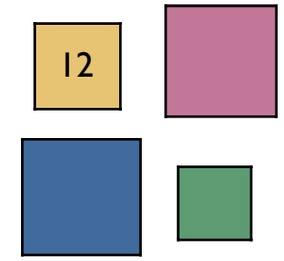
MANET



DataDesk



Mondrian



Selection-Sequences

- Why do we need Selection-Sequences?
  - A selection usually only exists as a set of selected points
    ⇒ no formal description of this set
  - Setting up complex selection sets is hard
    ⇒ errors are fatal, i.e. can not be re-done.
  - Alteration of the selection set is usually impossible
    ⇒ the complete selection must be repeated

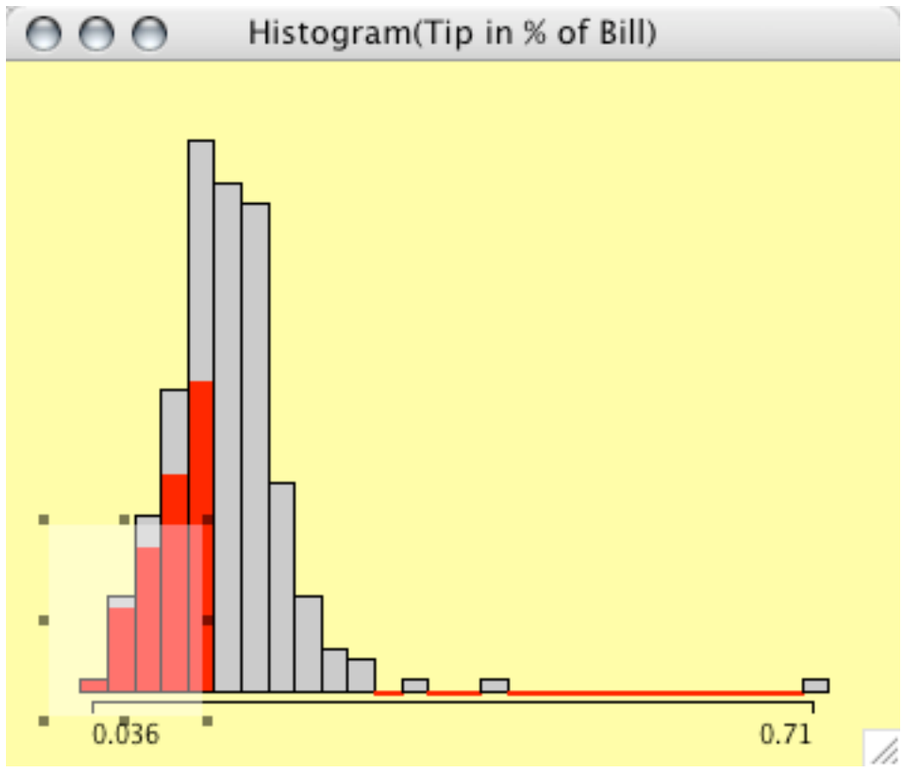
- Solution: Selection-Sequences
  - For each selection we store:
    - id
    - plot
    - coordinates
    - selection mode

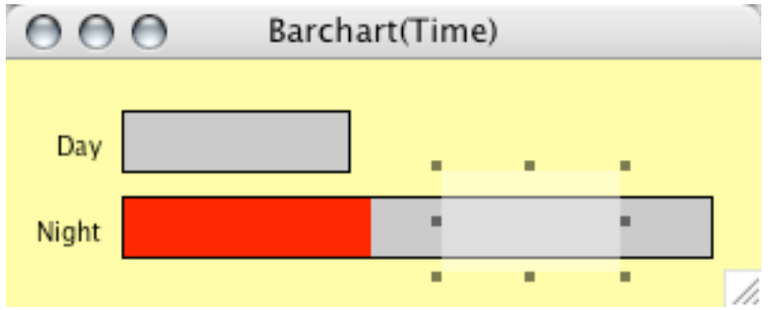
Selection-Sequences: Example

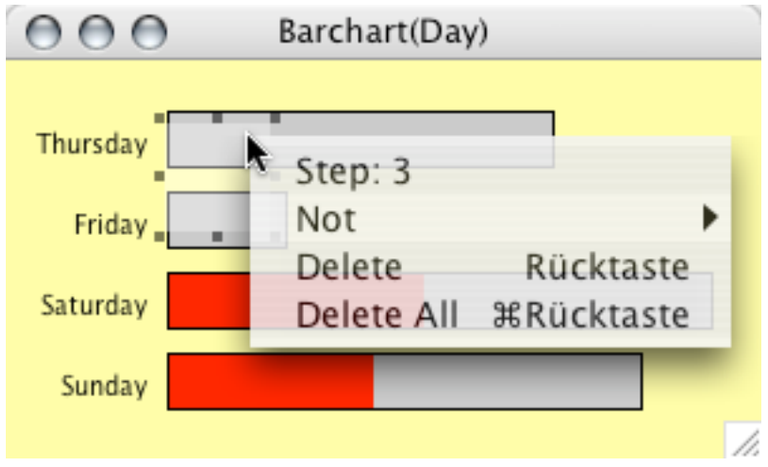
- Implementation in Mondrian

  Example (restaurant data):

  “Find all customers, who paid less than 15% tip, at night, except all days within the week!”







Selection-Sequences: Remarks

- Selection-Sequences are directed, i.e. for any three selection sets A, B, C

  A **OR** B **AND** C = A **OR** (B **AND** C) ≠ ((A **OR** B) **AND** C)

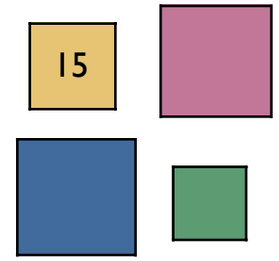
  and

  $A \cup B \cap C = A \cup (B \cap C) \neq ((A \cup B) \cap C)$

  holds, i.e. explicit left-parenthesis!

- Usually this is what the user was thinking about!

- Selection-Sequences can easily be translated into SQL. (Again, mind the left to right order of operators!)



Linking

- Linking between two instances (be it graphics, statistics or the data itself) of observations, can be categorized by their update mechanism:

  – **Cold**
  Cold linking is essentially NO linking, e.g. all graphics, summaries and data in **R** are cold linked.

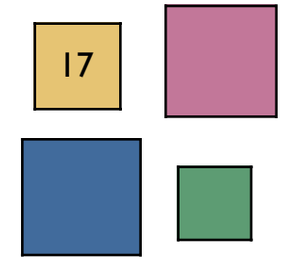
  – **Warm**
  If the user can decide whether or not a change in the data (selections, transformations, …) shall be propagated to a linked plot, we speak of a warm link

  – **Hot**
  Most interactive system support hot linking between all instances of the data, i.e. all changes propagate immediately to all other output windows.

Data Sources

- Basically two kinds of data sources can be distinguished
  - **Files**
    No matter what format the data is stored in a file (ASCII, Excel, R-Workspace), we usually assume, that the contents does not change, i.e. we are the only user accessing the file during a session.
  - **Databases**
    With databases, above concept no longer holds true.

    Issues with DB access
    - It is not advisable to keep a copy of the data in “private” memory
    - Data no longer can be assumed to be “constant”
    - Update strategies
      - Minimal: Manipulation in the software
      - Optional: Triggered from the DB (might be a bad idea!)
    - Direct DB access makes typical data mining far more efficient



Direct Manipulation

- **Direct** manipulation is at the core of any interactive system

  Typical manipulations comprise:

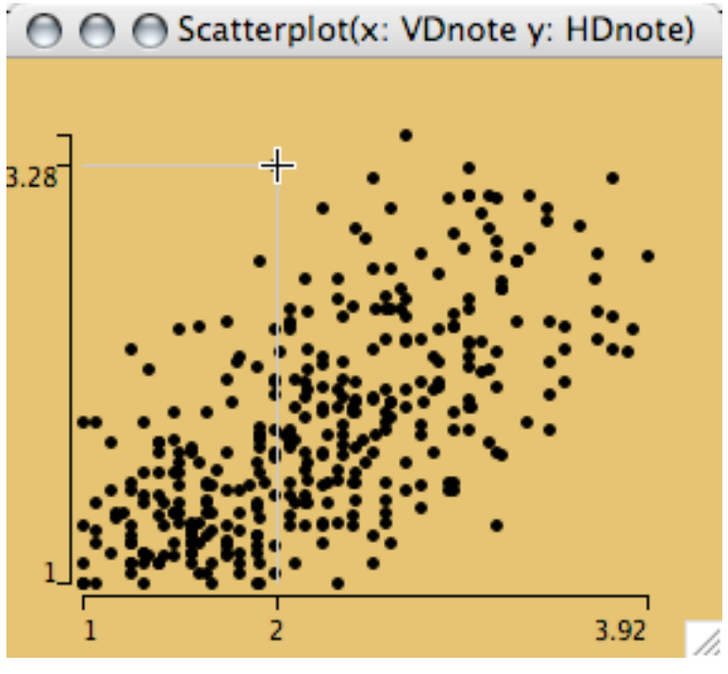
  - Selections ✔
  - Change of scale
    - axes
    - zoom
    - (order)
  - Change of order (both manually and automatically)
    - categories in a barchart
    - variables in parallel coordinates
    - variables in a mosaic plot
  - Change of plot parameters, e.g.
    - anchor point, bin width in histogram
    - point size in scatterplots
    - …

Queries

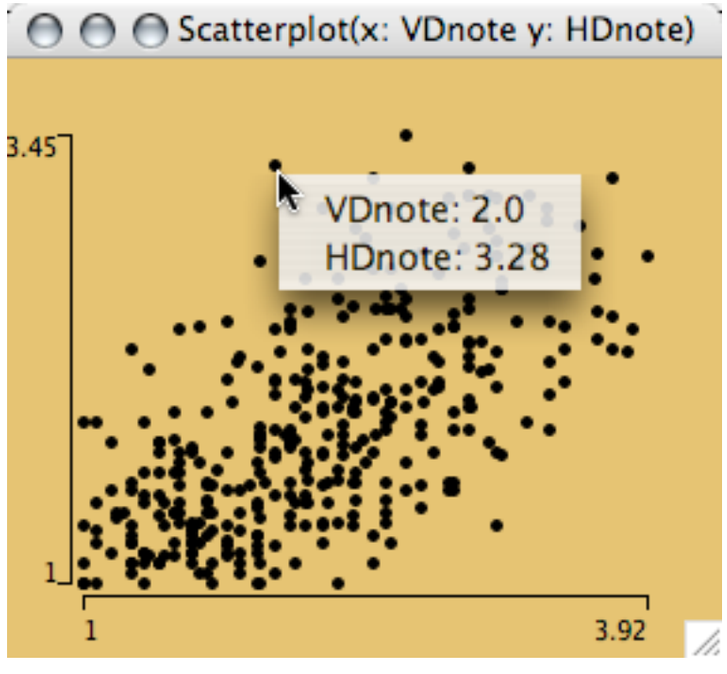
- Graphics are good at communicating qualitative information but fail to give exact quantities ⇒ need queries to get exact values
- Interactive graphics often display very little scale information (cf. Tufte's "data-ink-ratio")
- The level of detail of a query should have optional granularities,

e.g. scatterplot

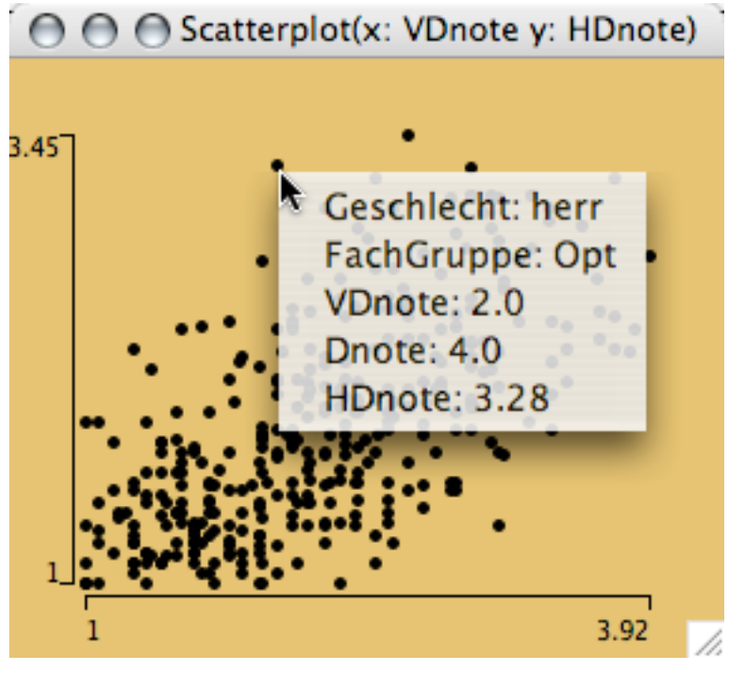
orientation



standard

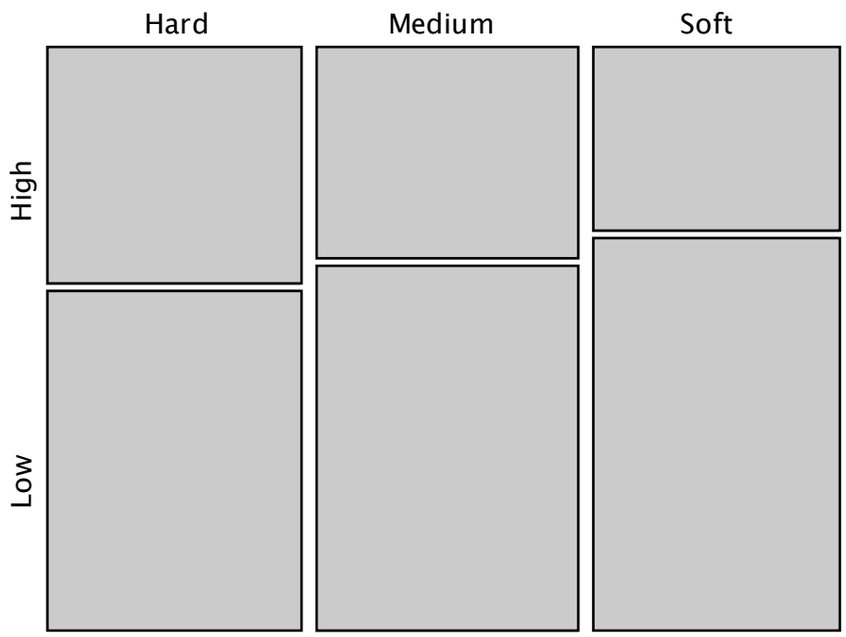


extended

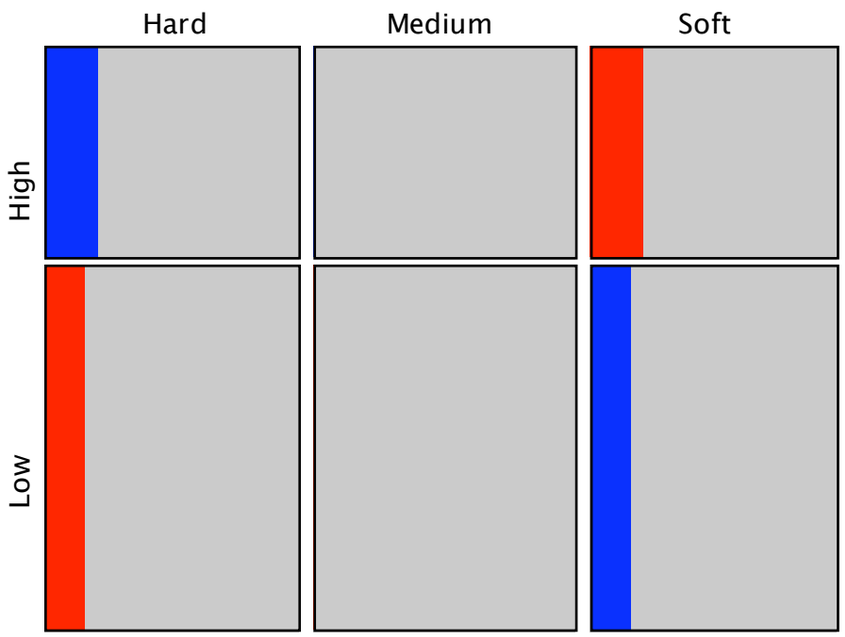


Enhanced Graphics

- Obviously, in interactive statistical graphics we are not only interested in whether “something is there”, but also whether “something is significant”
- Right now, only very little research has been done on enhancing graphics with “model information”
- Example:
  Loglinear models ⇔ Mosaic Plots



data ⇔ model

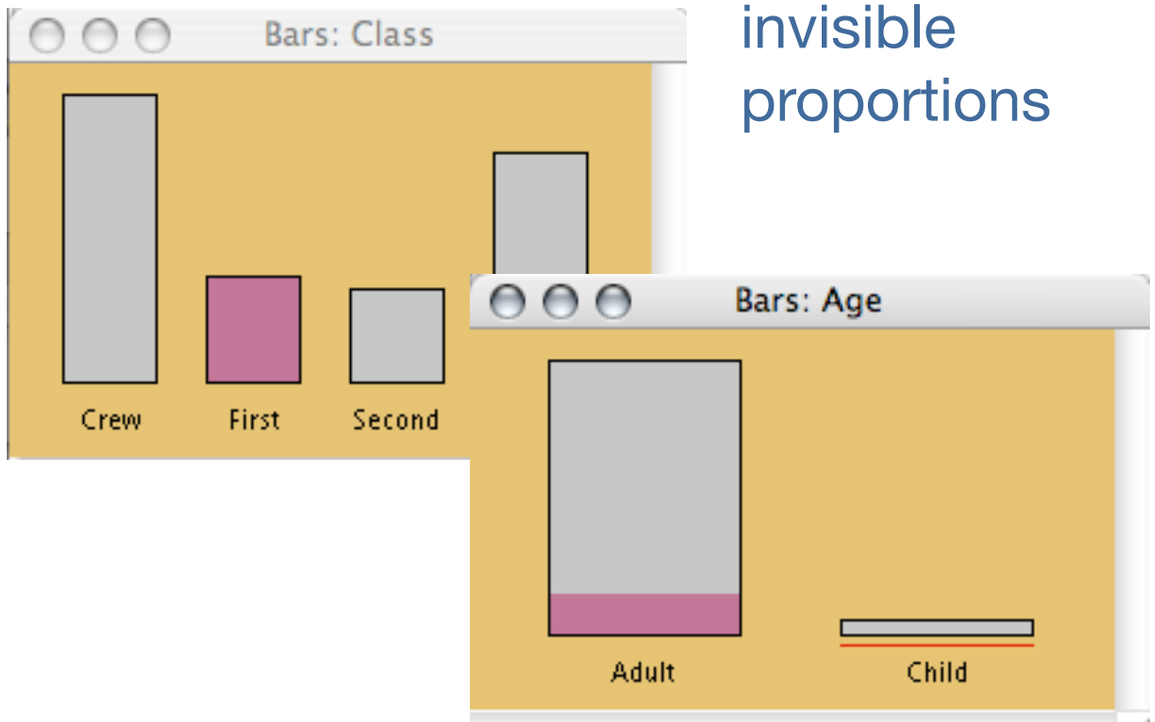


Warnings

- Two major points might be a cause of misperceptions in interactive systems:
  - Interactivity leads to rapid changes in data views and/or the data itself.
  - Limited screen resolution distorts the graphics and introduces rounding errors.

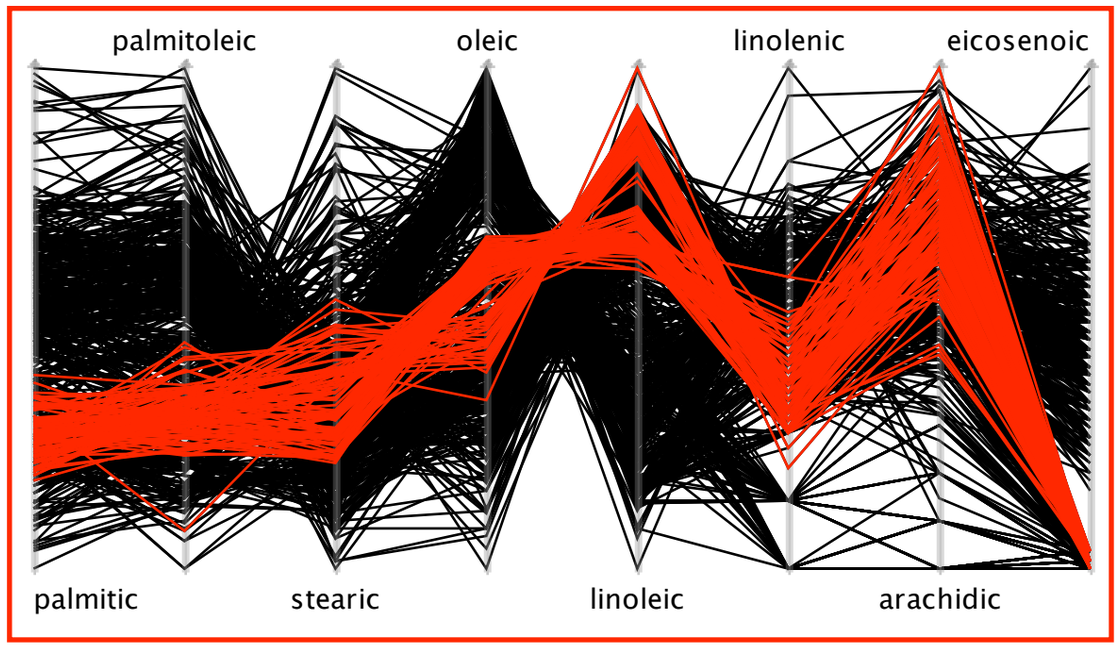
"What you see is not what you think you see!"

- Examples



invisible proportions

Outliers removed!



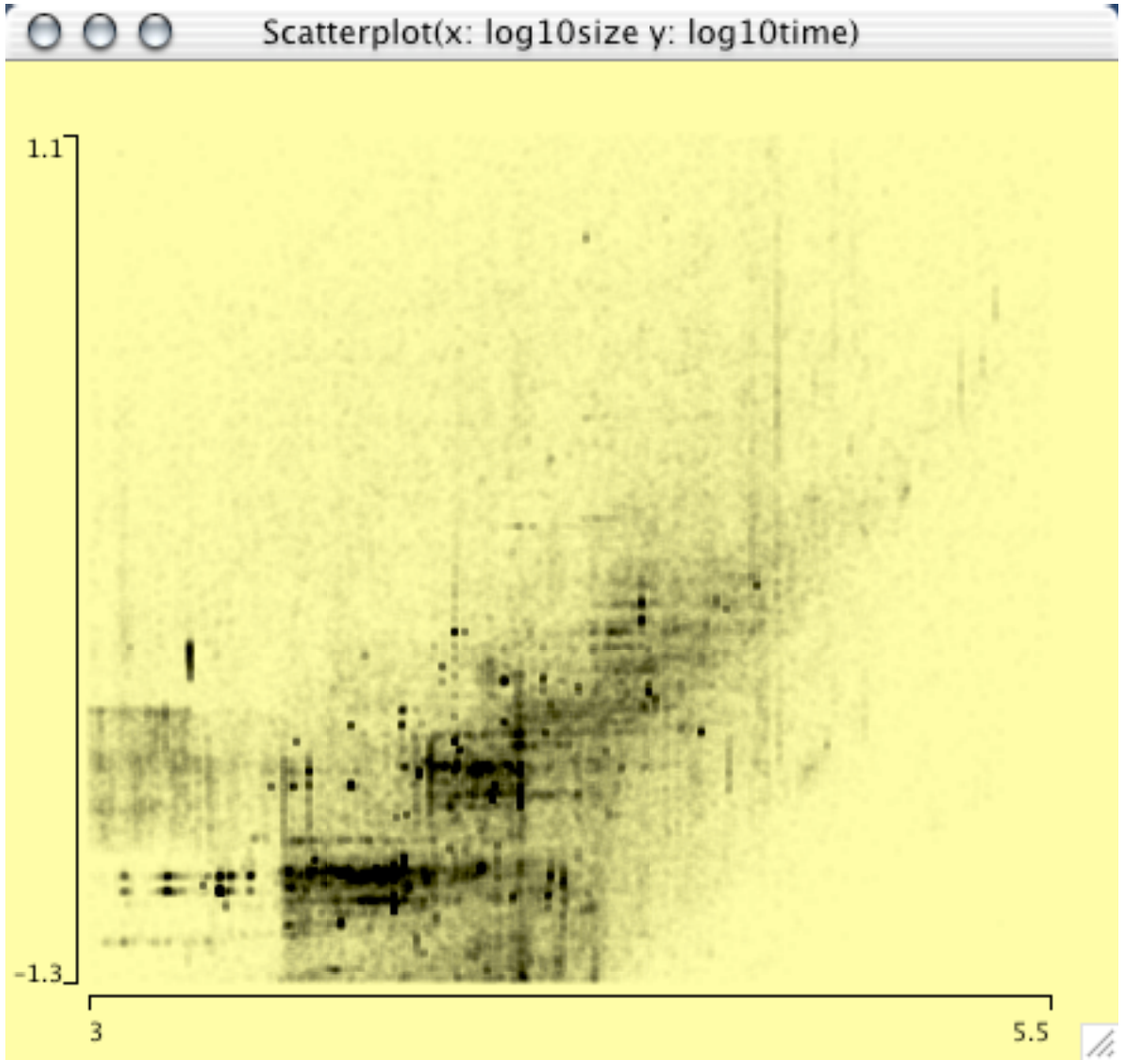
Rendering Issues

Two points are most important in rendering graphics on a computer screen:

- Rounding problems ✔
- Overplotting

  When dealing with really large data, overplotting can get very serious for all plots, that render a single glyph per observation.

α-transparency or pixel-binning can solve this problem.



382,127 internet transactions

Interactions: Conventions

- The usability of an interactive, computer based system can be strongly improved by strict conventions.

  This is a general concept for ALL user interfaces

- Typical interactions are:
  - selection
    - creation
    - change
  - reorder
  - queries
    - at different levels
  - zoom (**and logical zooming!**)
  - options

- All these actions MUST be implemented consistently across all plots and summaries

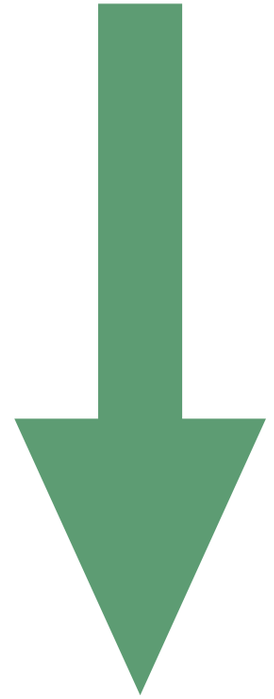
Interactions: Graphical User Interfaces (GUIs)

- Today's GUIs offer a sufficient variety of widgets to implement interactive statistical graphics software.

- Widget mapping:

Interaction	Widget	Action
Selection	Mouse-Click	Create Selection
Drag & Drop	Mouse + Modifier	Reorder, …
Cue	Cursor-Icon	DM of plot
Query	Tooltips	Query Info
Plot Alteration	Context Menu	Change Plot
Plot Alteration	Floating Palette	Change Plot
Plot Alteration	Modal Dialog Box	Change Plot

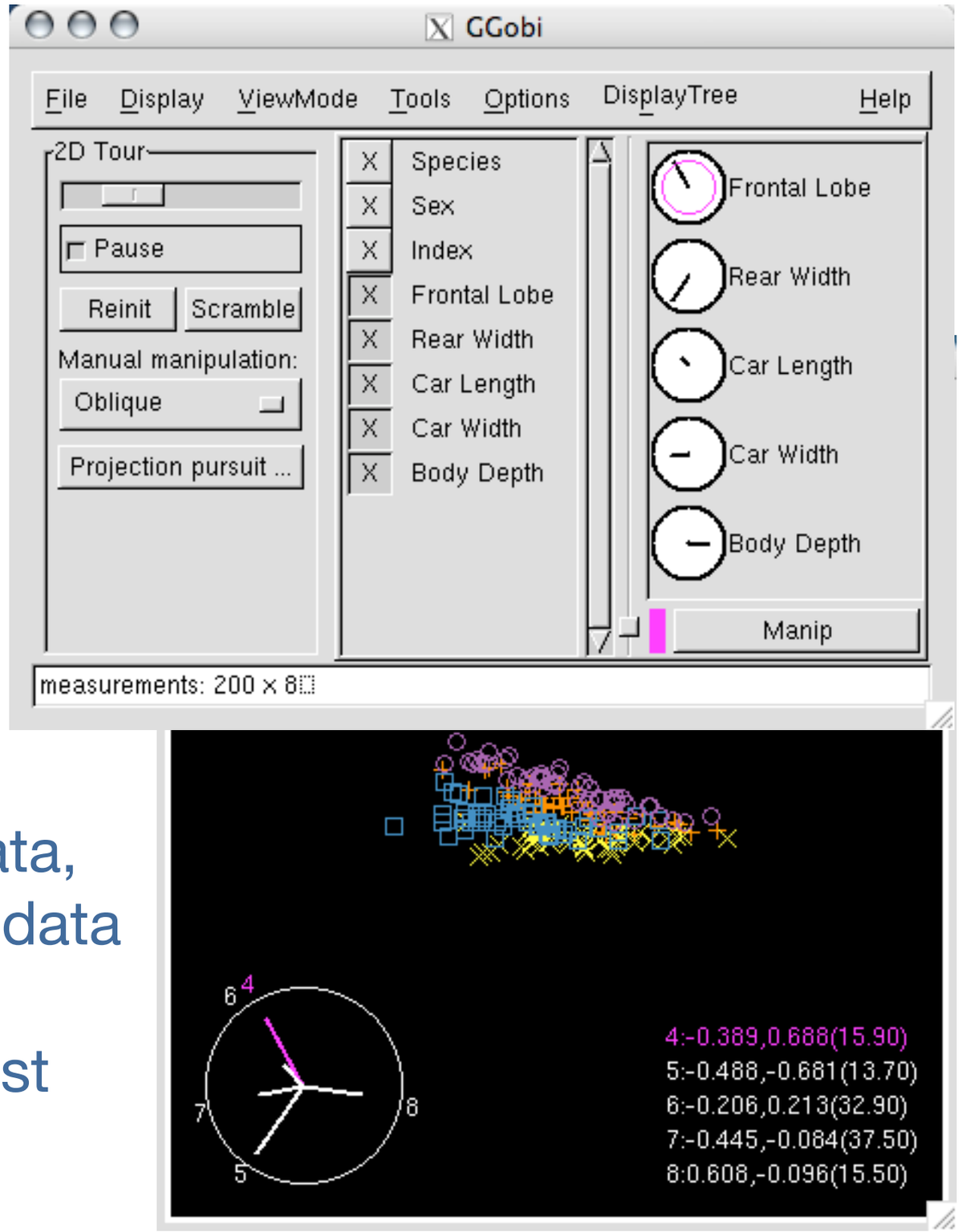
fast



slow

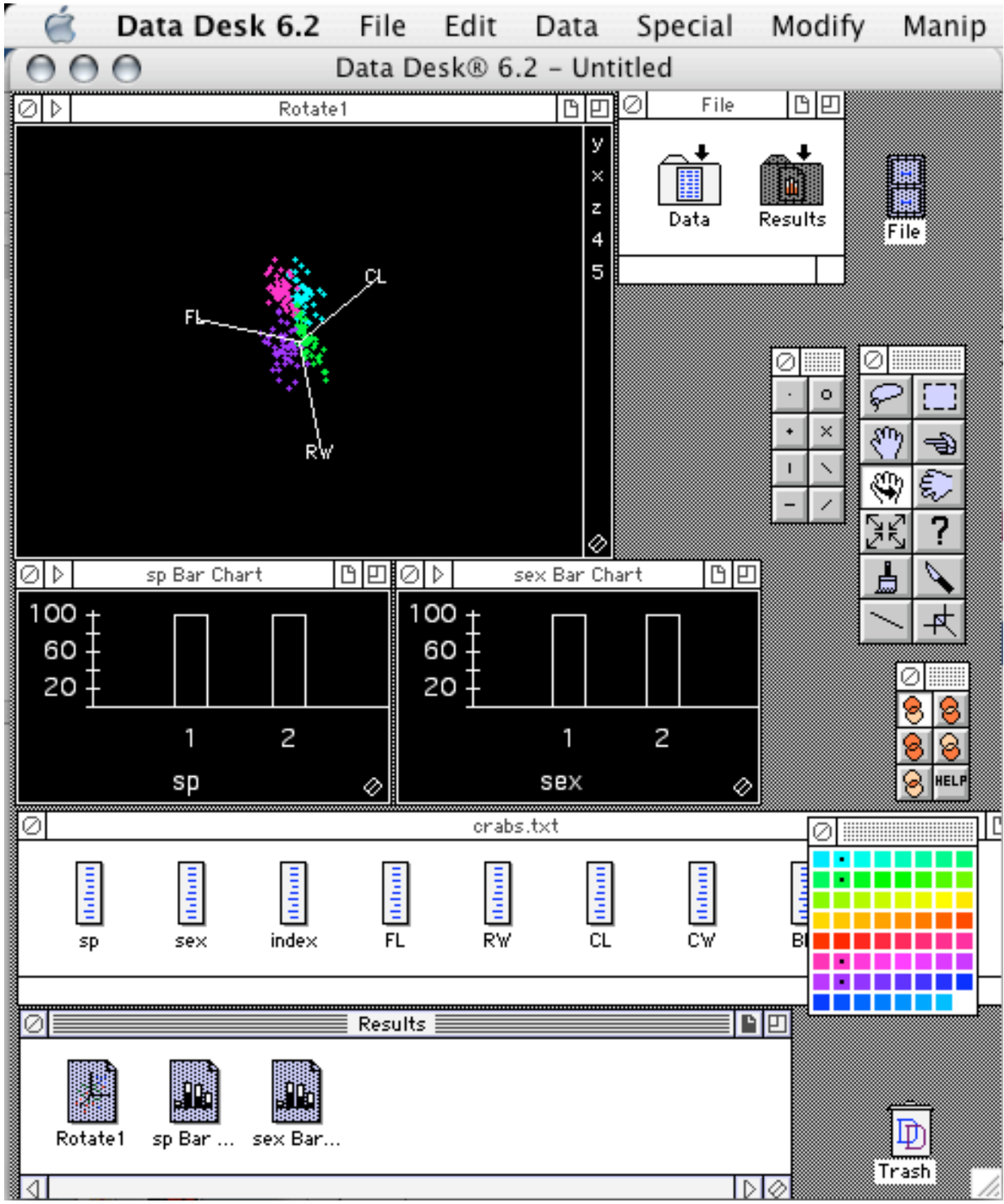
Examples: ggobi

- Initially started by Andreas Buja, Debby Swayne and Di Cook in 1990
- Implementation of Grand Tour
- Focus on high-dim scatterplots
- Still very “special” interface
- No support for non-numerical data, and little support for categorical data
- Runs on UNIXs and Windows (just about)



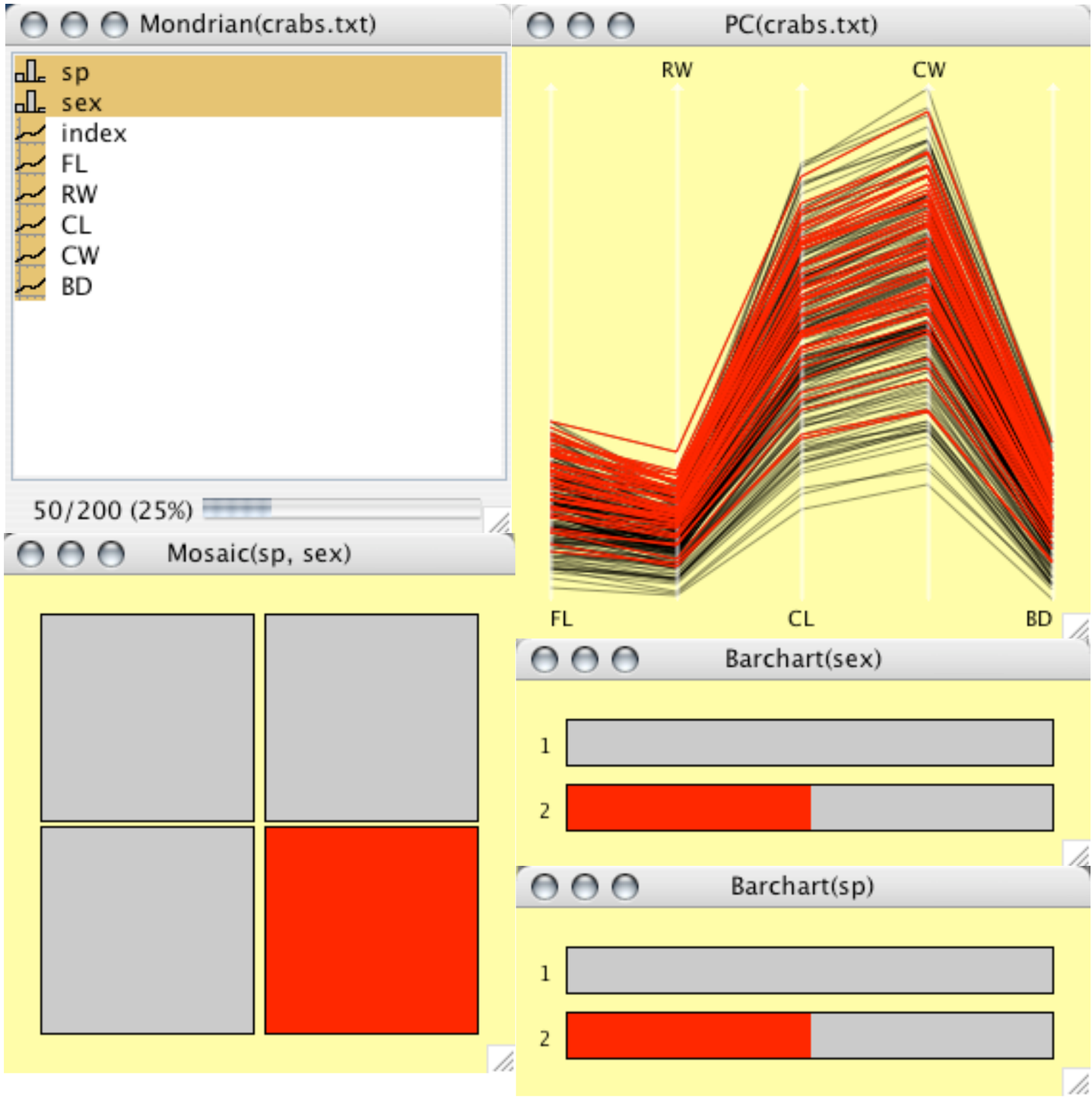
Examples: DataDesk

- First release in 1985 !
- Commercial package
- Combines graphical tools and statistical tools in **one** interactive environment
- Desktop and session concept
- No graphical support for high-dimensional data
- Development stalled since 1996



Examples: Mondrian

- Platform independent (JAVA)
- Focus on better support for categorical data
- Plots for high dimensional data and aware of large data sets
- Database connectivity only at experimental stage
- No desktop or sessions



"Shootout"

- Rough feature comparison of the three software tools

Feature	ggobi	DataDesk	Mondrian
Graphics	– (only tool for Grand Tour & Proj. Purs.)	– / o	+
Statistics	–	+	– / o
Desktop	–	+	–
Session	– / o	+	–
Meta Data	– / o	+	–
Selections	–	o	+
Queries	–	o	o / +
Speed	o / +	++	+
Interactivity	– / o	+	+

State of Play

- Interactive statistical graphics have developed substantially in the last decade
- Most (graphical) statistics software is not designed for interactivity
- Research software only delivers “proof of concept”, but not industrial strength
- Influence on commercial software is still small
- “Experts” need to use a variety of tools
- User interface standards are low – usually the interface is “in our way”
- Usability is crucial for user acceptance