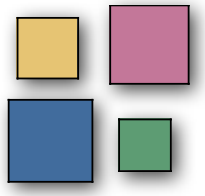


Interactive Graphics for Statistics

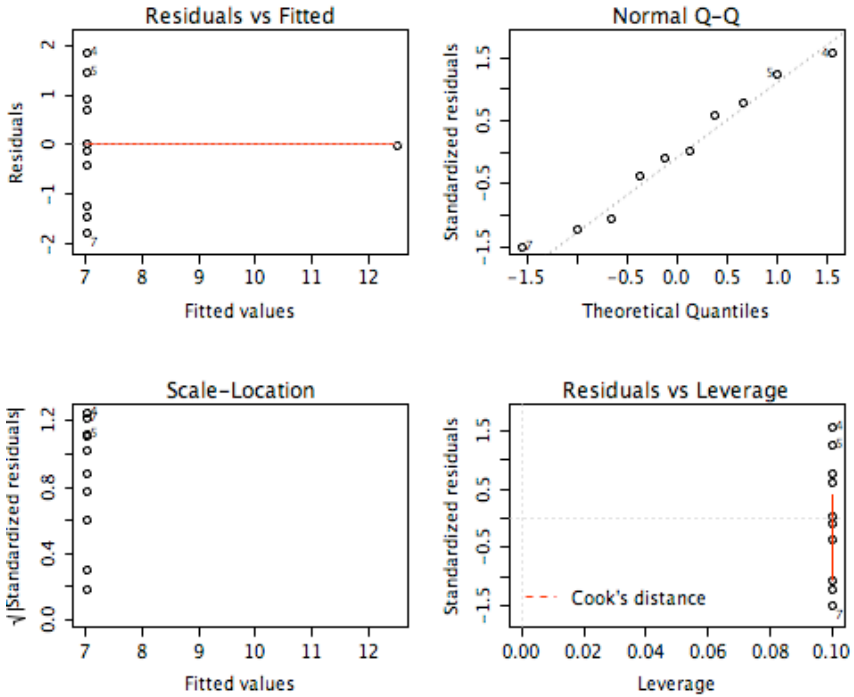
Principles & Examples

Martin Theus

martin.theus@math.uni-augsburg.de

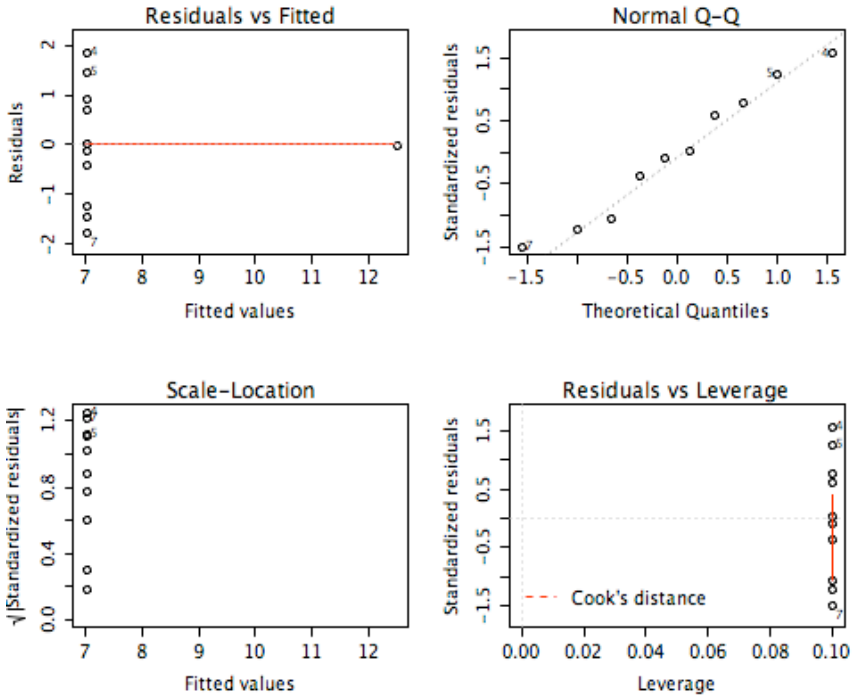
Graphics in Statistics: Diagnostics vs. Analysis

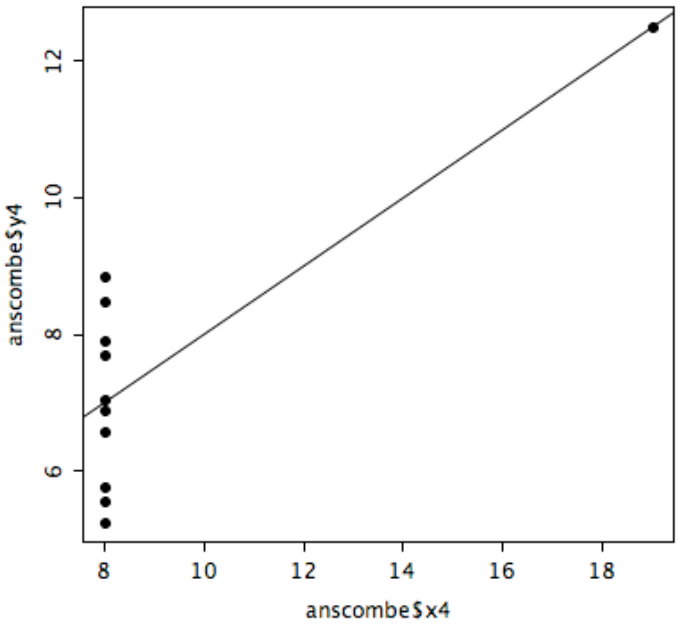
- Continuous Data …



Graphics in Statistics: Diagnostics vs. Analysis

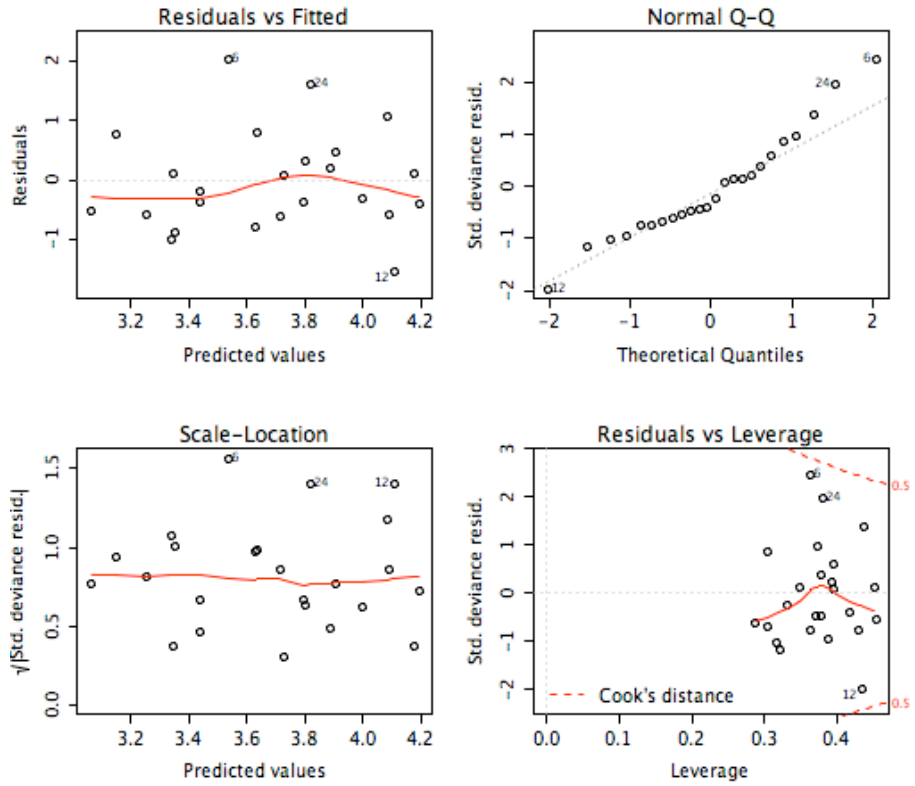
- Continuous Data …





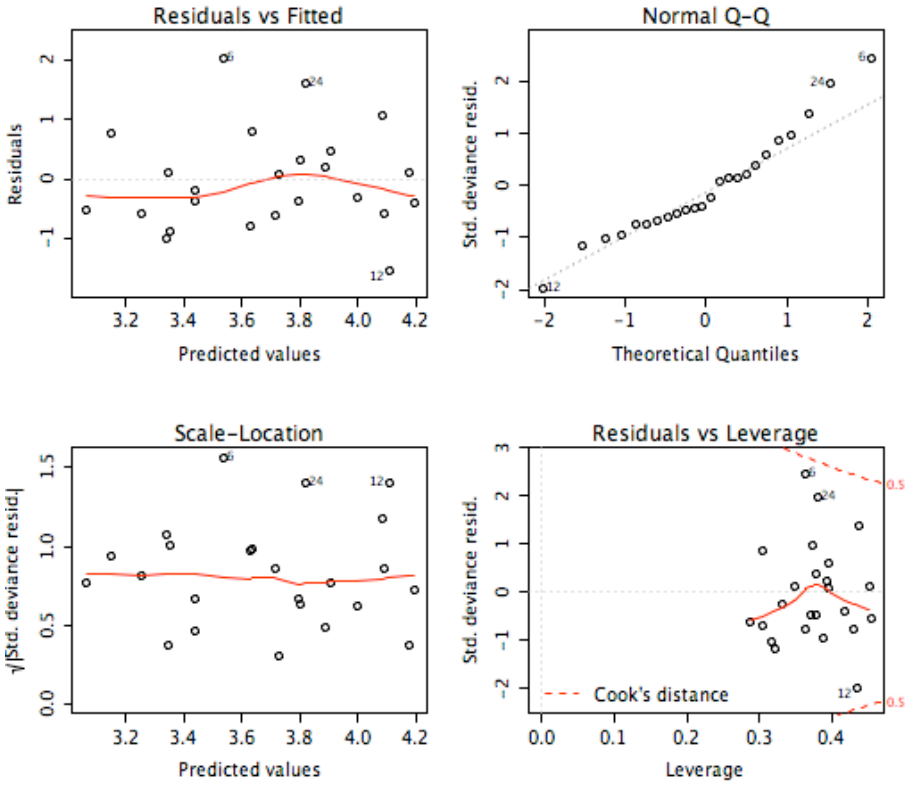
Graphics in Statistics: Diagnostics vs. Analysis

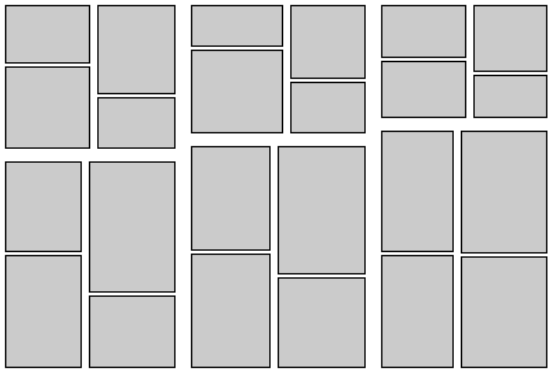
- Categorical Data …



Graphics in Statistics: Diagnostics vs. Analysis

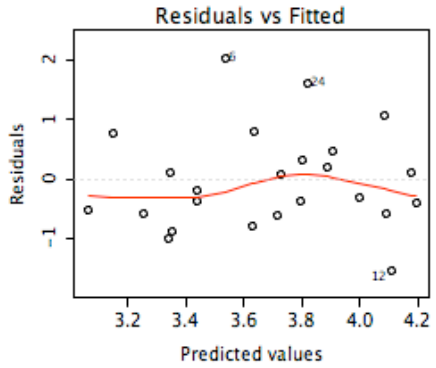
- Categorical Data …

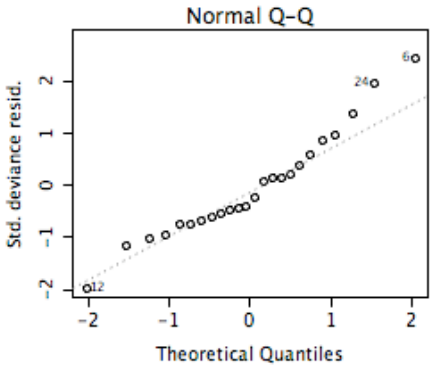


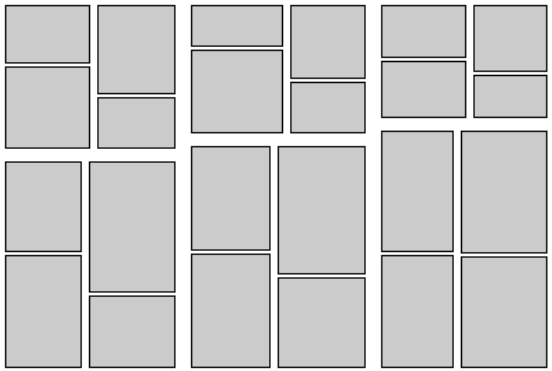


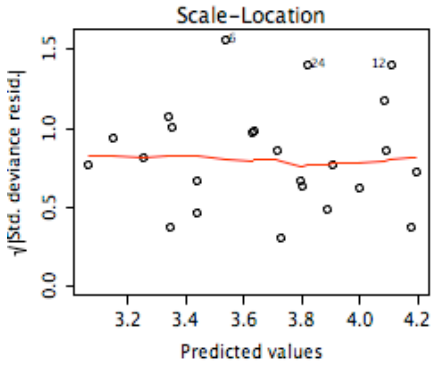
Graphics in Statistics: Diagnostics vs. Analysis

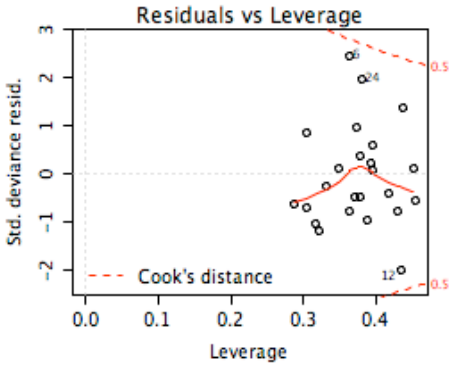
- Categorical Data …

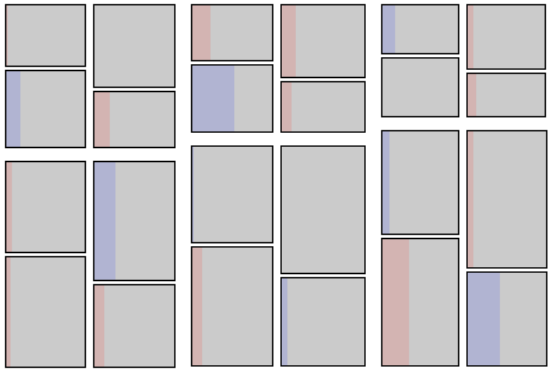


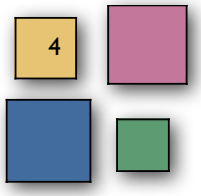












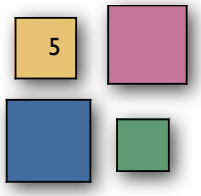
Graphics for Data Analysis ...

- is not based on formal theory
- but has proven to be very effective

- is not taught in standard curricula
- but used by most statisticians once they left college

- is supporting an interactive and iterative exploration of data
- thus needs interactive software tools

- is not well supported by most statistics software
- but there is software that helps



Interactive Graphical Tools for DA: History

Interactive Graphical Tools for DA: History

1973
PRIM-9
Tukey et al.

Interactive Graphical Tools for DA: History

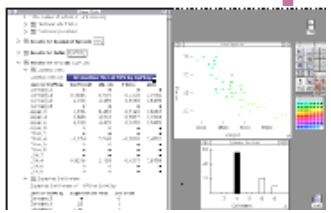
1983
SPLOM
Becker et al.

1973
PRIM-9
Tukey et al.

Interactive Graphical Tools for DA: History

1983
SPLOM
Becker et al.

1973
PRIM-9
Tukey et al.



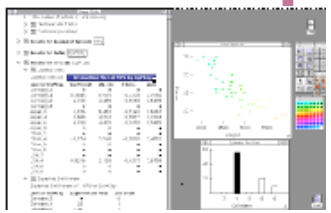
1985
DataDesk
Vellemann

Interactive Graphical Tools for DA: History

1983
SPLOM
Becker et al.

1991
XGobi
Buja et al.

1973
PRIM-9
Tukey et al.

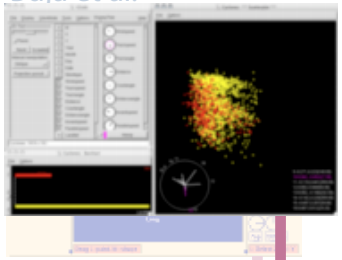


1985
DataDesk
Vellemann

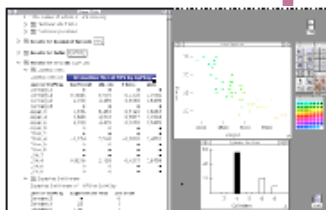
Interactive Graphical Tools for DA: History

1983
SPLOM
Becker et al.

1999
ggobi
Swayne et al.
Buja et al.



1973
PRIM-9
Tukey et al.

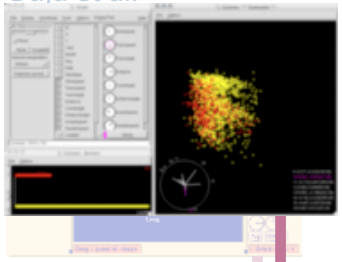


1985
DataDesk
Vellemann

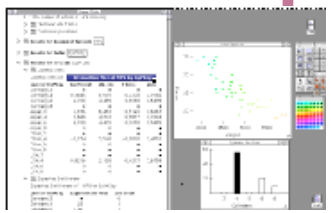
Interactive Graphical Tools for DA: History

1983
SPLOM
Becker et al.

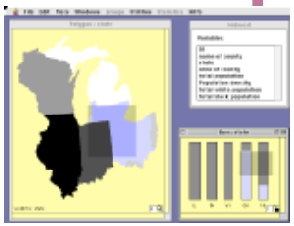
1999
ggobi
Swayne et al.
Buja et al.



1973
PRIM-9
Tukey et al.



1985
DataDesk
Vellemann

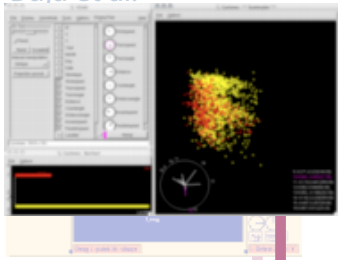


1993
MANET
Unwin et al.

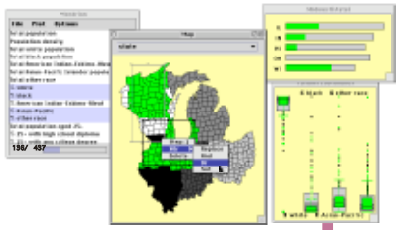
Interactive Graphical Tools for DA: History

1983
SPLOM
Becker et al.

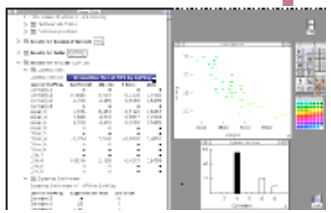
1999
ggobi
Swayne et al.
Buja et al.



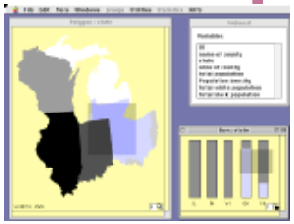
1997
Mondrian
Theus



1973
PRIM-9
Tukey et al.



1985
DataDesk
Vellemann

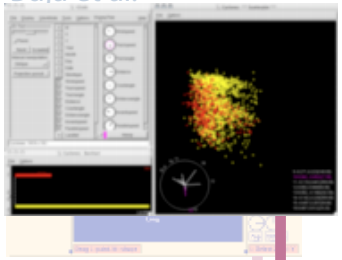


1993
MANET
Unwin et al.

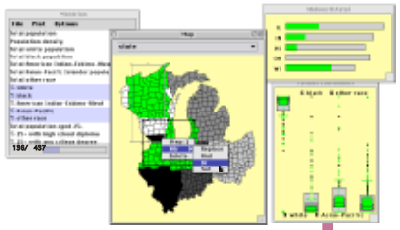
Interactive Graphical Tools for DA: History

1983
SPLOM
Becker et al.

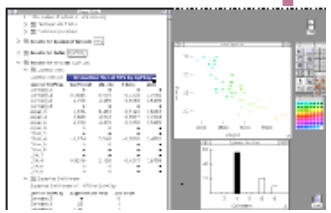
1999
ggobi
Swayne et al.
Buja et al.



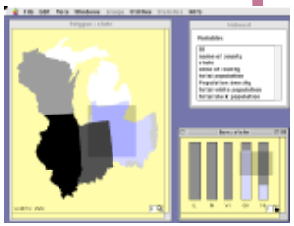
1997
Mondrian
Theus



1973
PRIM-9
Tukey et al.



1985
DataDesk
Vellemann



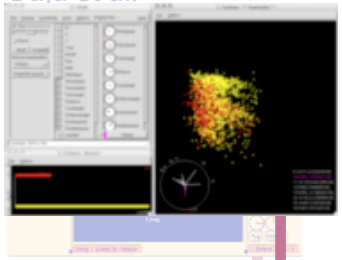
1993
MANET
Unwin et al.

2006
StatStudio
SAS

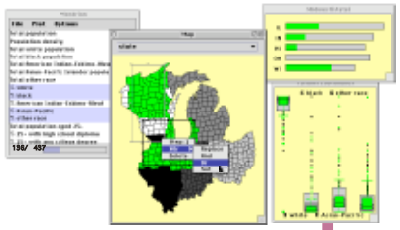
Interactive Graphical Tools for DA: History

1983
SPLOM
Becker et al.

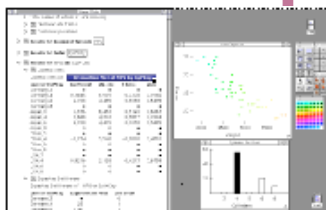
1999
ggobi
Swayne et al.



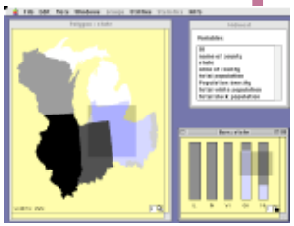
1997
Mondrian
Theus



19
PRIM-9
Tukey et al.



1985
DataDesk
Vellemann



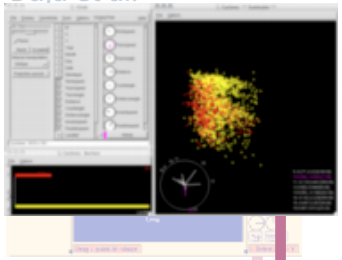
1993
MANET
Unwin et al.

2006
StatStudio
SAS

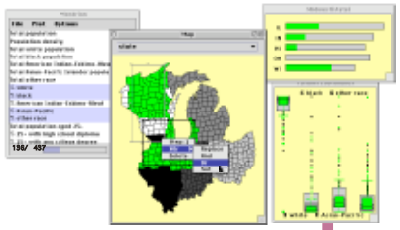
Interactive Graphical Tools for DA: History

1983
SPLOM
Becker et al.

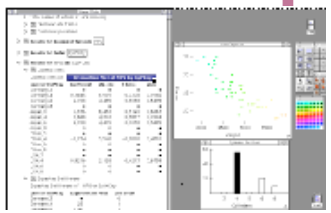
1999
ggobi
Swayne et al.
Buja et al.



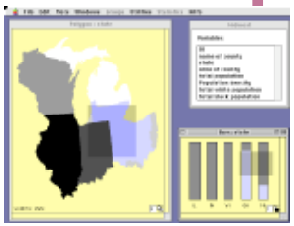
1997
Mondrian
Theus



19
PRIM-9
Tukey et al.



1985
DataDesk
Vellemann



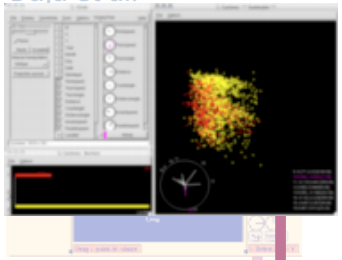
1993
MANET
Unwin et al.

2006
StatStudio
SAS

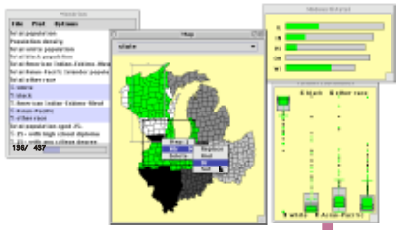
Interactive Graphical Tools for DA: History

1983
SPLOM
Becker et al.

1999
ggobi
Swayne et al.
Buja et al.

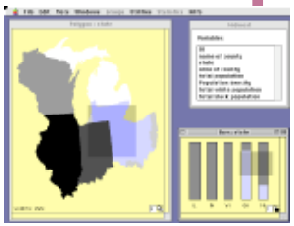


1997
Mondrian
Theus



19
PRIM-9
Tukey et al.

1985
DataDesk
Vellemann



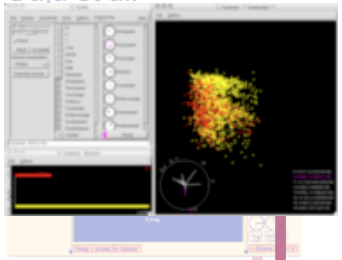
1993
MANET
Unwin et al.

2006
StatStudio
SAS

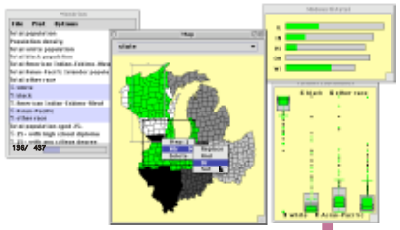
Interactive Graphical Tools for DA: History

1983
SPLOM
Becker et al.

1999
ggobi
Swayne et al.
Buja et al.

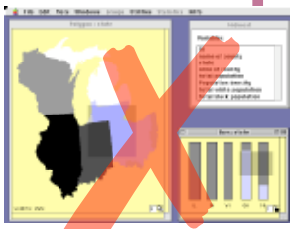


1997
Mondrian
Theus



19
PRIM-9
Tukey et al.

1985
DataDesk
Vellemann



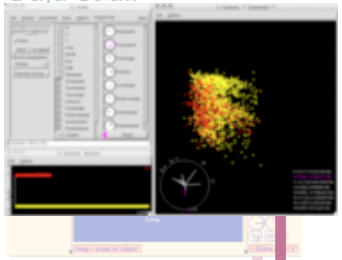
1993
MANET
Unwin et al.

2006
StatStudio
SAS

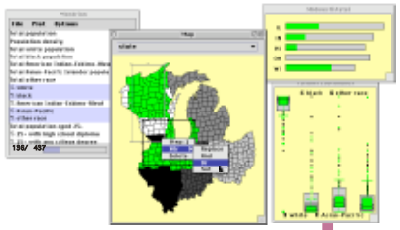
Interactive Graphical Tools for DA: History

1983
SPLOM
Becker et al.

1999
ggobi
Swayne et al.
Buja et al.

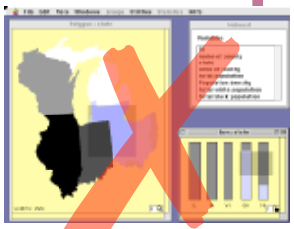


1997
Mondrian
Theus



19
PRIM-9
Tukey et al.

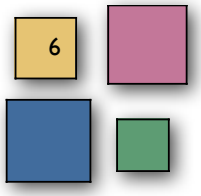
1985
DataDesk
Vellemann



1993
MANET
Unwin et al.

2006
StatStudio
SAS

people forget far faster than they learn ⇒ give them time!



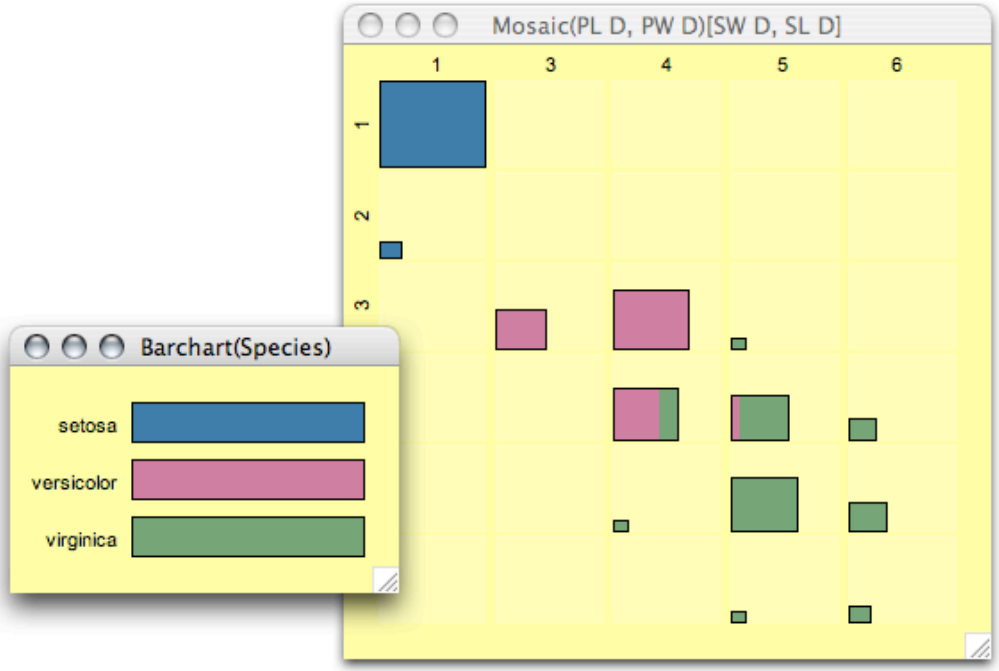
On Hammers and Nails ...

- No one single tool is universal enough to cover every problem we might want to handle in data analysis ⇒ use the best suited!?

On Hammers and Nails ...

- No one single tool is universal enough to cover every problem we might want to handle in data analysis ⇒ use the best suited!?

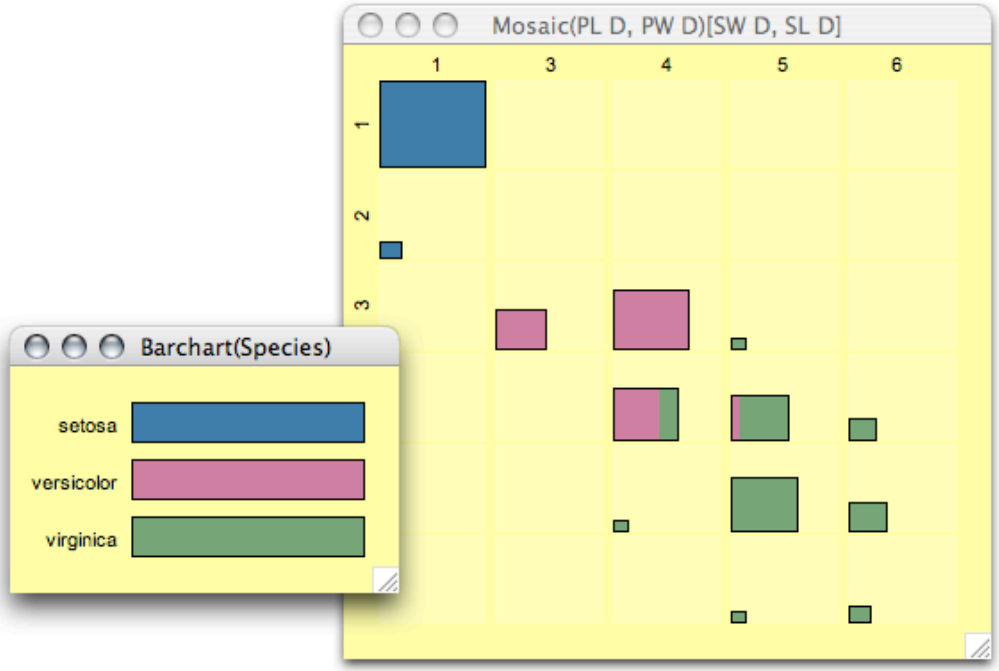
Iris data in a Fluctuation Diagram



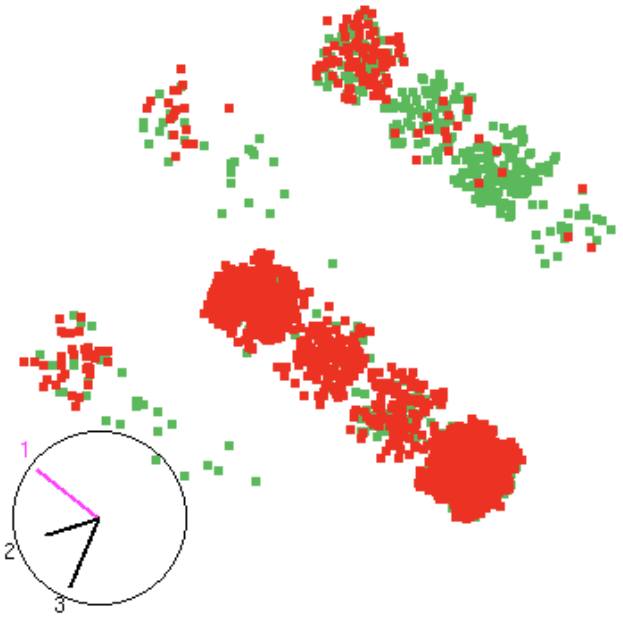
On Hammers and Nails ...

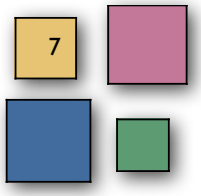
- No one single tool is universal enough to cover every problem we might want to handle in data analysis ⇒ use the best suited!?

Iris data in a Fluctuation Diagram

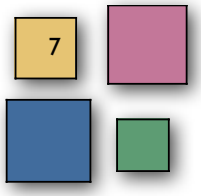


Titanic data in a scatterplot





What is Interactive Statistical Graphics about?

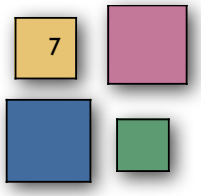


What is Interactive Statistical Graphics about?

- **Direct manipulation is at the core**

  (classical computational statistics was 90% algorithm and 10% interface – Interactive Statistical Graphics is 10% algorithm and 90% interface!)

  - Selection of subgroups, i.e., conditioning
  - Modification of plot parameters



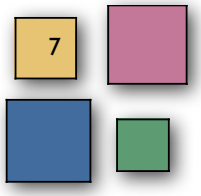
What is Interactive Statistical Graphics about?

- **Direct manipulation is at the core**

  (classical computational statistics was 90% algorithm and 10% interface – Interactive Statistical Graphics is 10% algorithm and 90% interface!)

  - Selection of subgroups, i.e., conditioning
  - Modification of plot parameters

- **What makes a graphics interactive?**
  - data can be **selected**
    (selection state is actually only one possible attribute of the data)
  - support for **highlighting**
  - objects can be **queried**



What is Interactive Statistical Graphics about?

- **Direct manipulation is at the core**

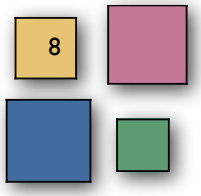
  (classical computational statistics was 90% algorithm and 10% interface – Interactive Statistical Graphics is 10% algorithm and 90% interface!)

  - Selection of subgroups, i.e., conditioning
  - Modification of plot parameters

- **What makes a graphics interactive?**
  - data can be **selected**
    (selection state is actually only one possible attribute of the data)
  - support for **highlighting**
  - objects can be **queried**

- **Interactive Graphics $\neq$ Dynamic Graphics**

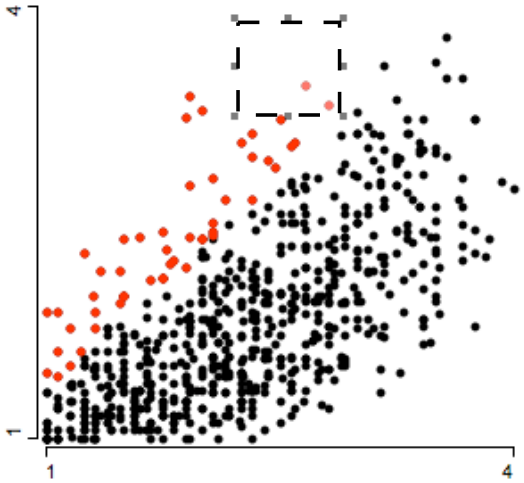
  dynamic variation of plot parameters (animation) was the first interaction to be implemented 30 years ago, but is now out-dated.



Where did Linking go?

- Linking is used to (optionally) propagate attributes of the data

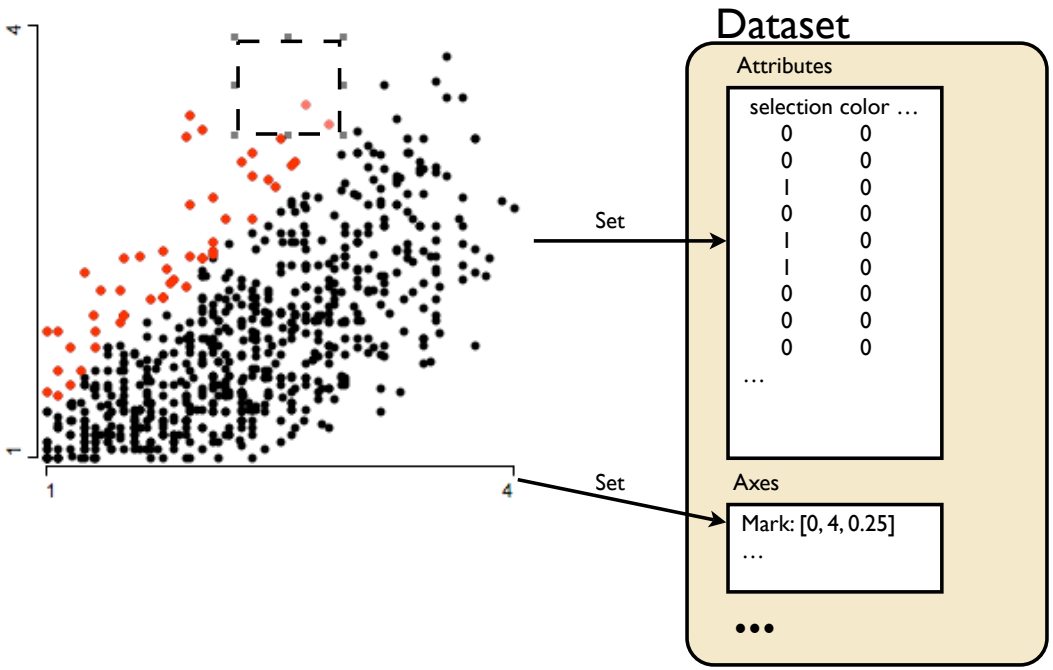
Example:



Where did Linking go?

- Linking is used to (optionally) propagate attributes of the data

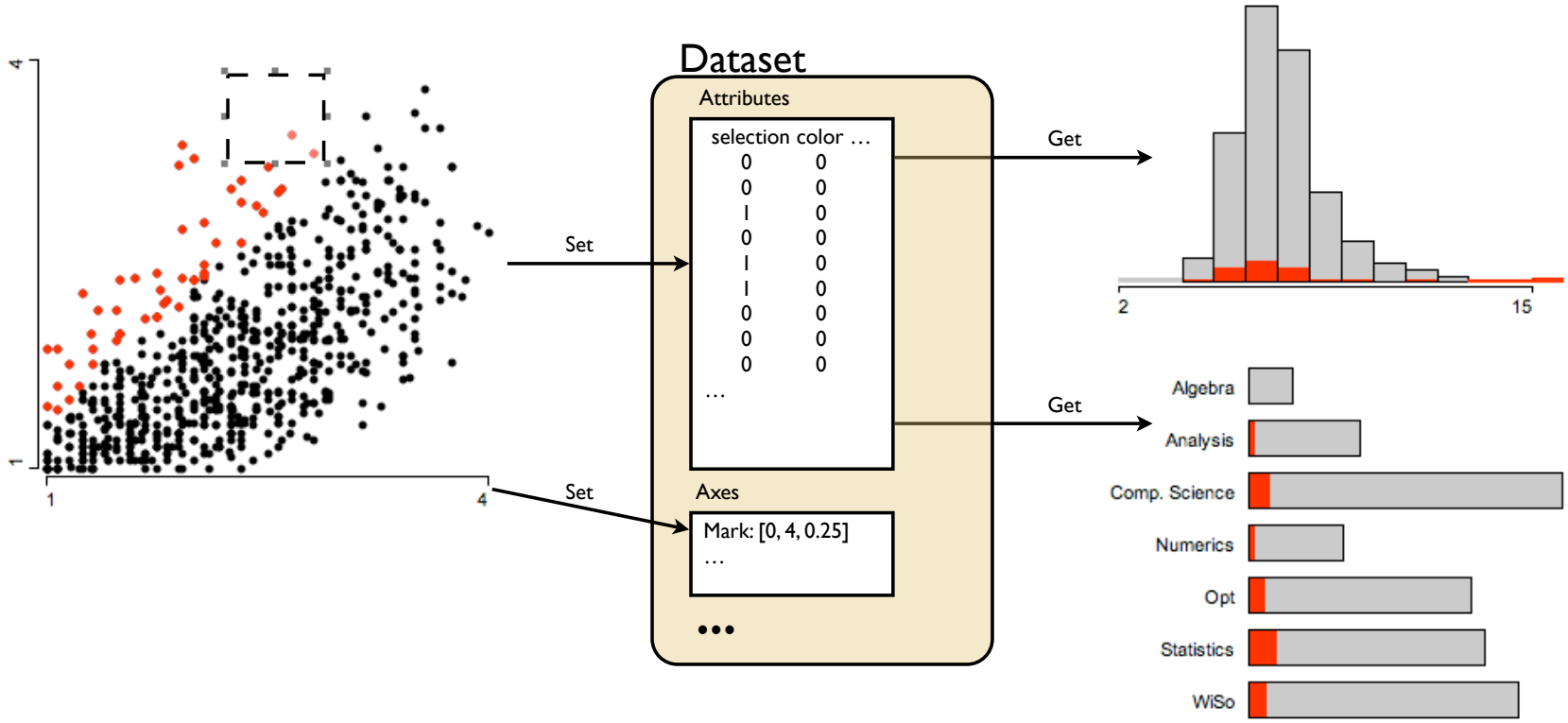
Example:

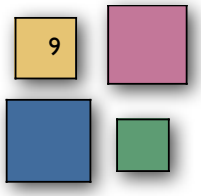


Where did Linking go?

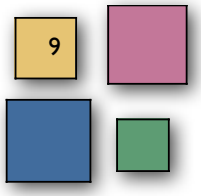
- Linking is used to (optionally) propagate attributes of the data

Example:



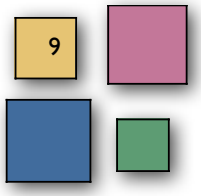


More on Linking ...



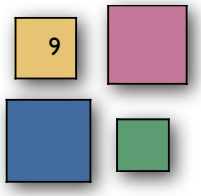
More on Linking ...

- Linking is not restricted to the highlight state



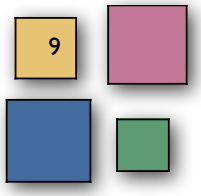
More on Linking ...

- Linking is not restricted to the highlight state
- Case linking
  - (transient) highlighting
  - (persistent) color brushing
  - symbols
  - visibility



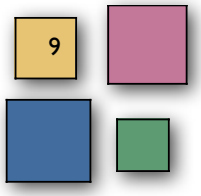
More on Linking ...

- Linking is not restricted to the highlight state
- Case linking
  - (transient) highlighting
  - (persistent) color brushing
  - symbols
  - visibility
- Axes linking
  - continuous intervals
  - categorical orderings



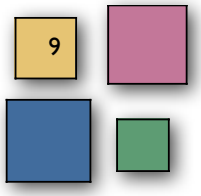
More on Linking ...

- Linking is not restricted to the highlight state
- Case linking
  - (transient) highlighting
  - (persistent) color brushing
  - symbols
  - visibility
- Axes linking
  - continuous intervals
  - categorical orderings
- Parameter linking



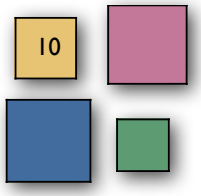
More on Linking ...

- Linking is not restricted to the highlight state
- Case linking
  - (transient) highlighting
  - (persistent) color brushing
  - symbols
  - visibility
- Axes linking
  - continuous intervals
  - categorical orderings
- Parameter linking
- Linking of attributes of the data (or any other “incarnation” of the data) is **optionally**.



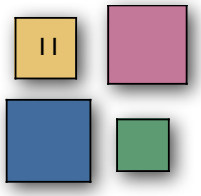
More on Linking ...

- Linking is not restricted to the highlight state
- Case linking
  - (transient) highlighting
  - (persistent) color brushing
  - symbols
  - visibility
- Axes linking
  - continuous intervals
  - categorical orderings
- Parameter linking
- Linking of attributes of the data (or any other “incarnation” of the data) is **optionally**.
- Setting and/or getting of attributes may have **constraints** (group selections, cross-relation selections, …)



More on Selections …

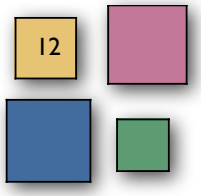
- Different tools can be used (point, rectangle, slice, lasso, brush, …) – most of them select parallel or orthogonal to variable axes.
- Successive selection steps must be combinable with boolean operators (single step selections can only select trivial sets).
- Multiple brushes can be used to define a sequence of selections.
- Painting (persistent brush selection) is actually covered by a selection sequence when all brushes are in “OR” mode.
- Prefer parametric selection over arbitrary selections.
  (How to communicate a lasso selection, or a selection in some high-dimensional rotated space?)



More on Highlighting ...

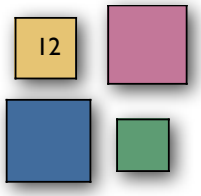
- Implementation of highlighting for glyph-based plots is trivial.
- Highlighting for rectangle based plots is still straight forward.
- The highlighting in a plot not necessarily is of the same type as the plot itself, but it must be faithful to the plots definition.
- Highlighting is a transient attribute of the data and usually compares to the complete sample,

  Color brushing defines a persistent group membership, and thus the groups should compare between each other.
- Not all plots can be used to implement (sensible) highlighting …



More on Queries ...

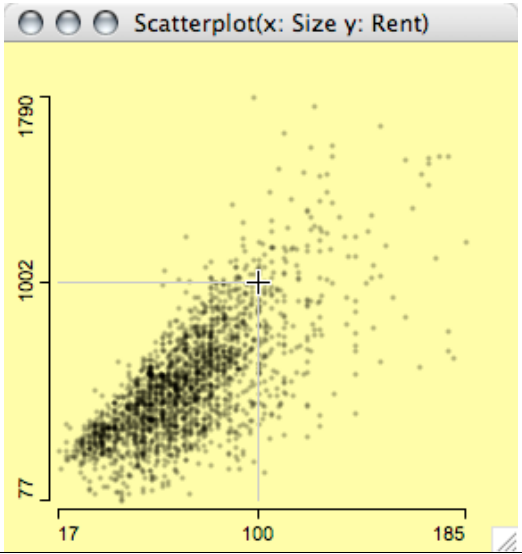
- Graphics are good at communicating qualitative information but fail to give exact quantities ⇒ need queries to get exact values.
- Interactive graphics often display very little scale information (cf. Tufte's "data-ink-ratio").
- The level of detail of a query should have optional granularities, e.g.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



More on 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:

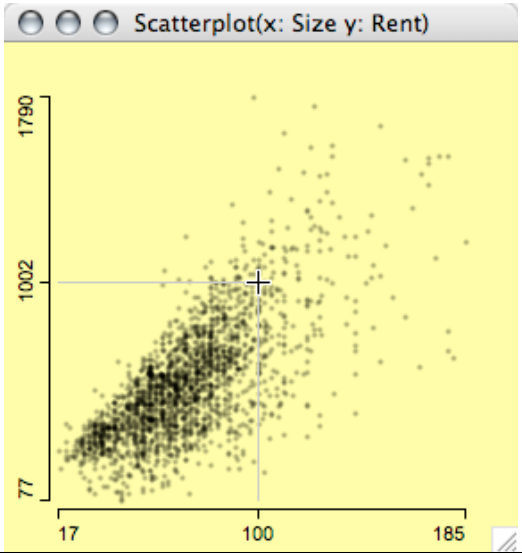
orientation



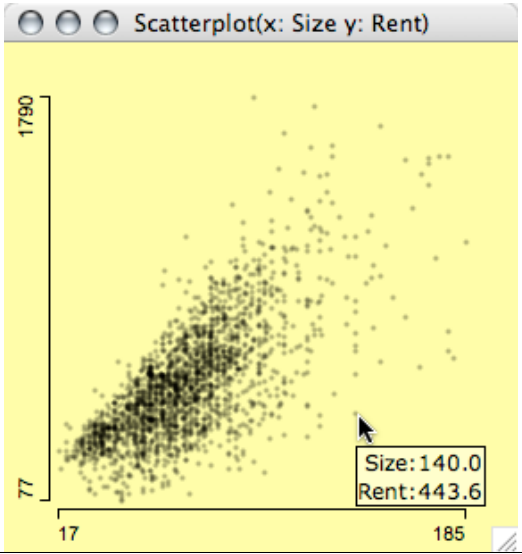
More on Queries ...

- 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



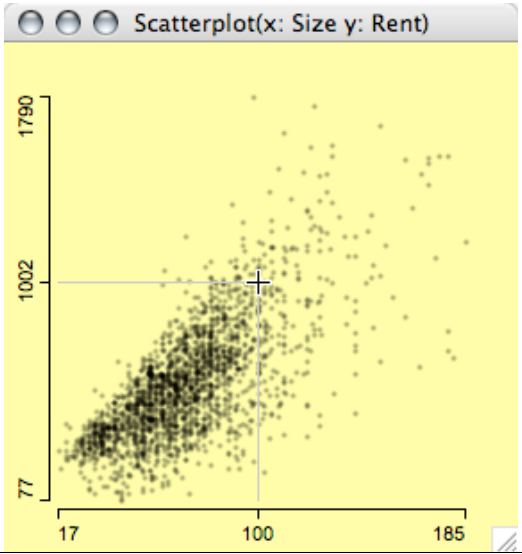
standard



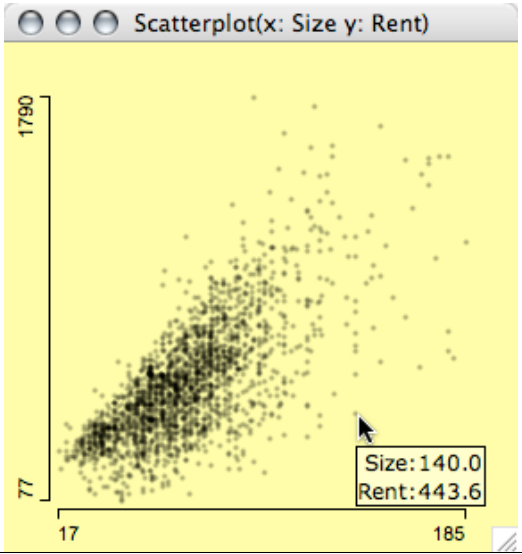
More on Queries ...

- 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:

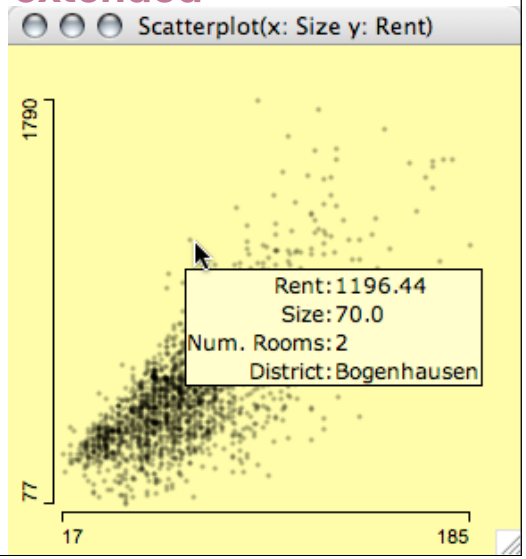
orientation

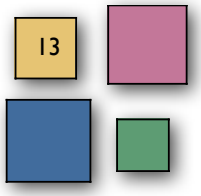


standard



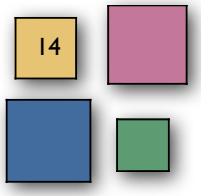
extended





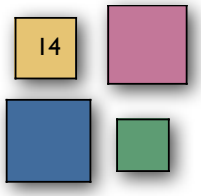
Interacting with Plots

- Typical manipulations comprise of:
  - Selections ✔
  - Change of scale
    - axes
    - zoom (includes logical zooming)
    - (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
    - …
- Important: Identify general concepts!



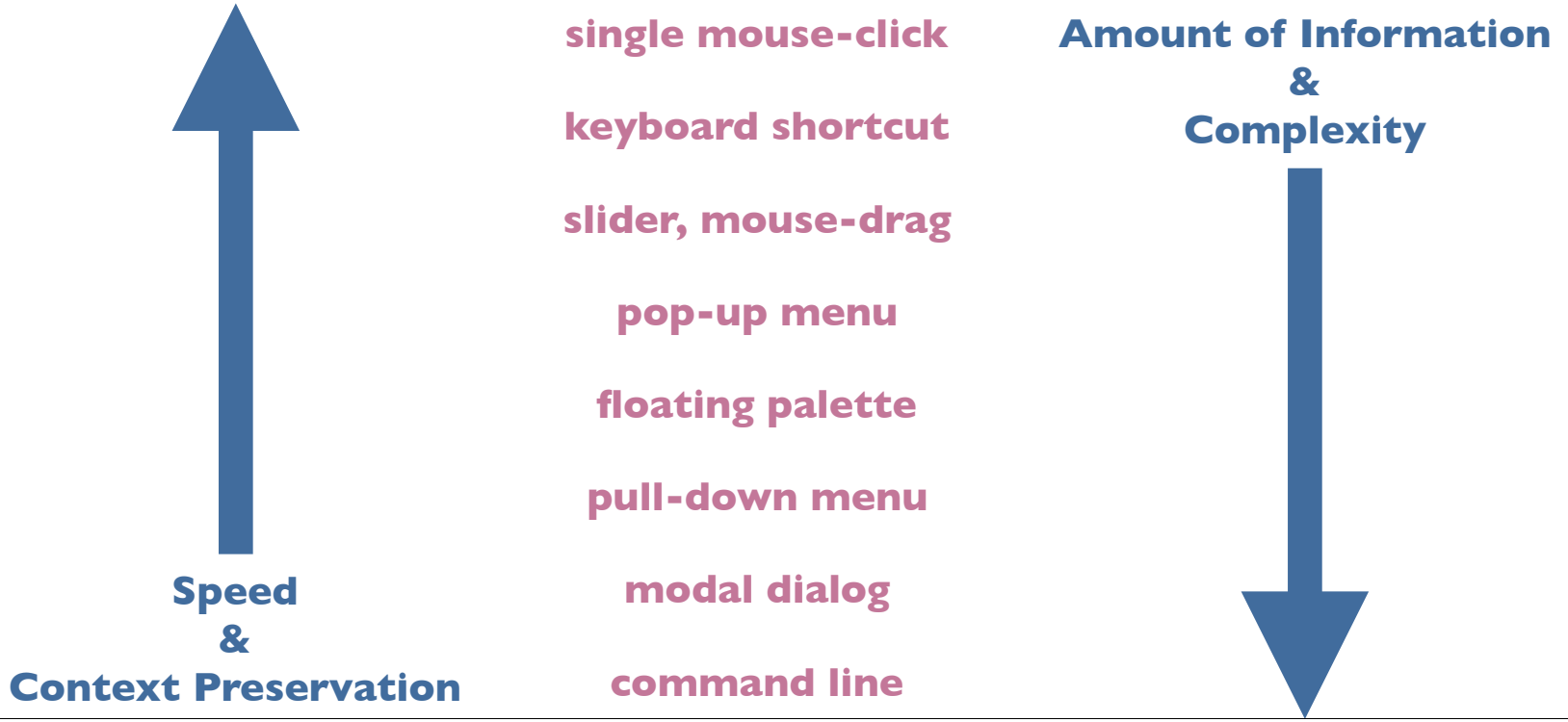
On the Interface

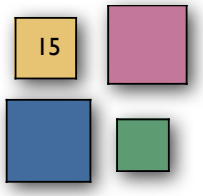
- The usability of an interactive, computer based system can be strongly improved by strict conventions.
- All interactions (selection, queries, ordering, zoom, …) **MUST** be implemented consistently across all plots and summaries.



On the Interface

- The usability of an interactive, computer based system can be strongly improved by strict conventions.
- All interactions (selection, queries, ordering, zoom, …) **MUST** be implemented consistently across all plots and summaries.

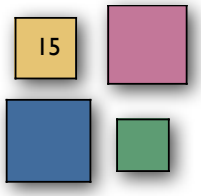




Less can be more ...

- Software should offer a clearly defined (orthogonal) set of functions, which can be combined to get a wider range of functionality.

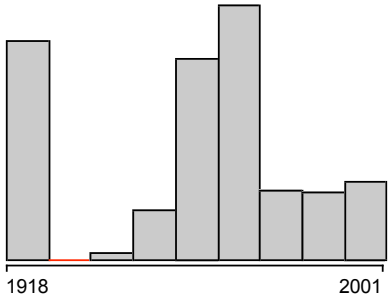
- Examples:
  - Painting Mode is a brush in OR mode and covered by selection sequences ✓
  - Categorizing a continuous variable via a histogram

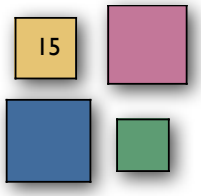


Less can be more ...

- Software should offer a clearly defined (orthogonal) set of functions, which can be combined to get a wider range of functionality.

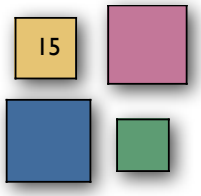
- Examples:
  - Painting Mode is a brush in OR mode and covered by selection sequences ✓
  - Categorizing a continuous variable via a histogram





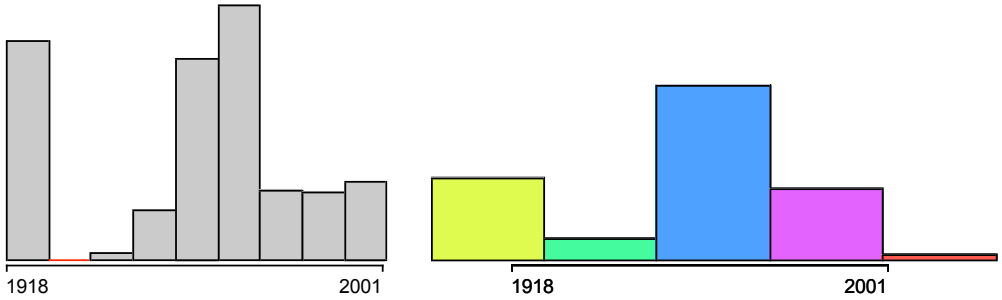
Less can be more ...

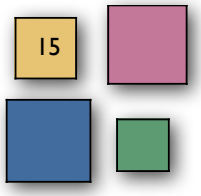
- Software should offer a clearly defined (orthogonal) set of functions, which can be combined to get a wider range of functionality.
- Examples:
  - Painting Mode is a brush in OR mode and covered by selection sequences ✓
  - Categorizing a continuous variable via a histogram



Less can be more ...

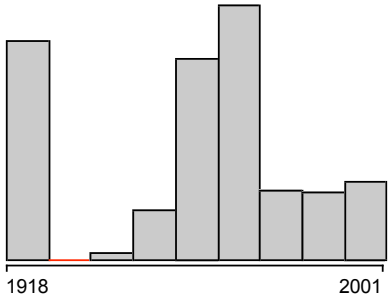
- Software should offer a clearly defined (orthogonal) set of functions, which can be combined to get a wider range of functionality.
- Examples:
  - Painting Mode is a brush in OR mode and covered by selection sequences ✓
  - Categorizing a continuous variable via a histogram

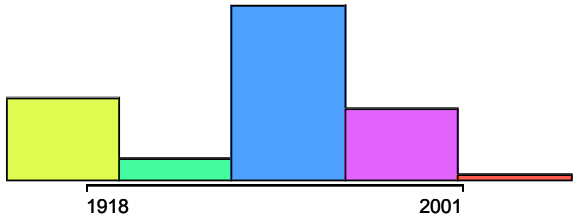


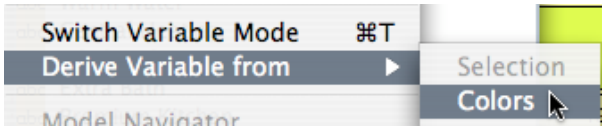


Less can be more ...

- Software should offer a clearly defined (orthogonal) set of functions, which can be combined to get a wider range of functionality.
- Examples:
  - Painting Mode is a brush in OR mode and covered by selection sequences ✓
  - Categorizing a continuous variable via a histogram

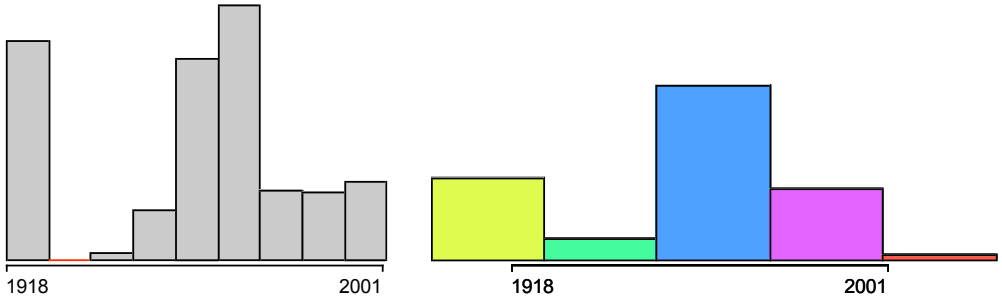


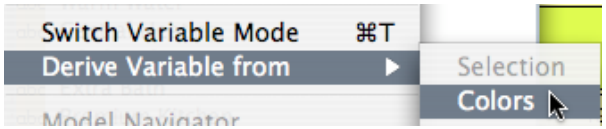


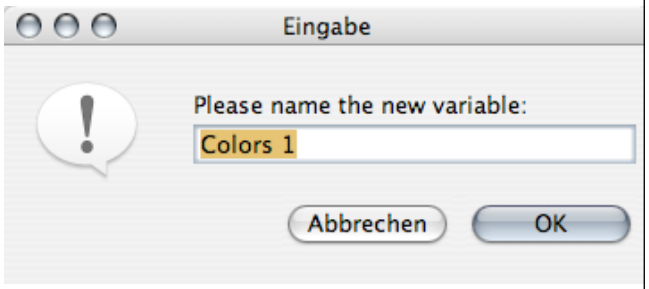


Less can be more ...

- Software should offer a clearly defined (orthogonal) set of functions, which can be combined to get a wider range of functionality.
- Examples:
  - Painting Mode is a brush in OR mode and covered by selection sequences ✓
  - Categorizing a continuous variable via a histogram

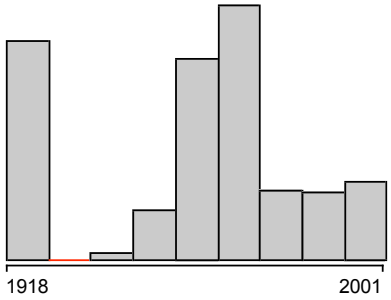


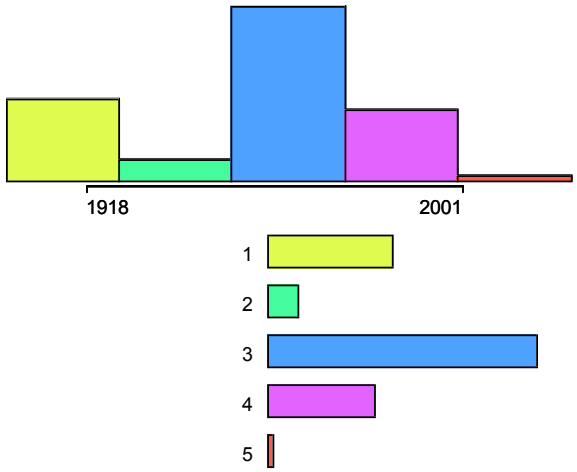


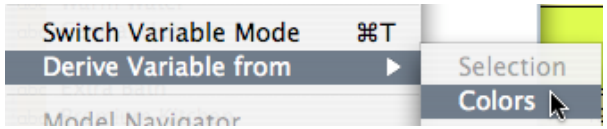


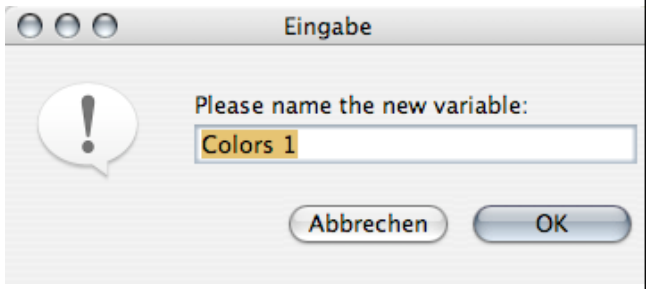
Less can be more ...

- Software should offer a clearly defined (orthogonal) set of functions, which can be combined to get a wider range of functionality.
- Examples:
  - Painting Mode is a brush in OR mode and covered by selection sequences ✓
  - Categorizing a continuous variable via a histogram



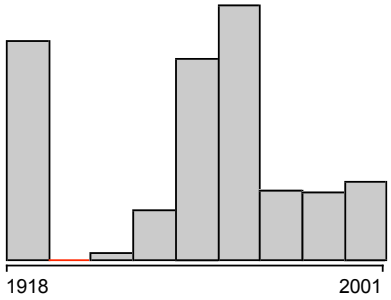


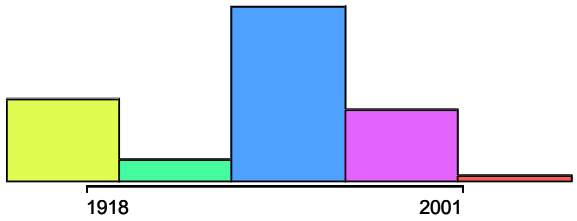


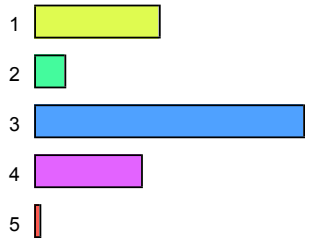


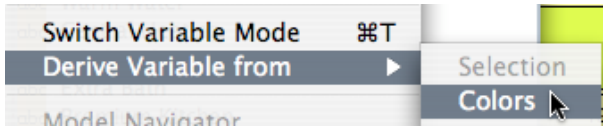
Less can be more ...

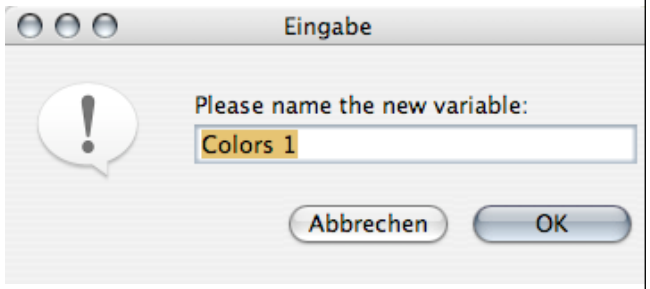
- Software should offer a clearly defined (orthogonal) set of functions, which can be combined to get a wider range of functionality.
- Examples:
  - Painting Mode is a brush in OR mode and covered by selection sequences ✓
  - Categorizing a continuous variable via a histogram



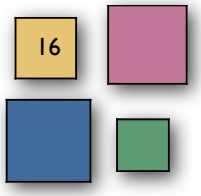








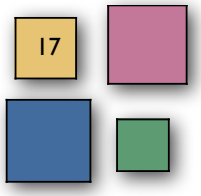
- Functions are less obvious for a novice user!



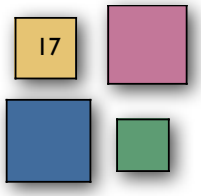
Rendering Quality is an Issue!

- Most graphics systems (S, R, SAS, …) were designed at times where “pen on paper” was the rendering model.
- Drawing on a computer screen will introduce rounding problem; the size of small objects is usually far off from what it should be.
- Anti-aliased graphics and sub-pixel rendering are not only a “matter of taste”.
- Overplotting is not an inherent problem of computer graphics, but if software does not handle it decently, the interpretation of (glyph based) graphics can lead to wrong conclusions.

  (extremely misleading with color brushing!)

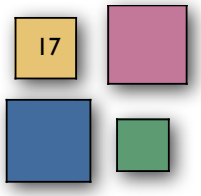


Conclusion



Conclusion

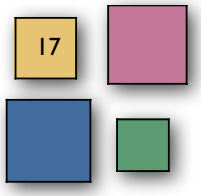
- Graphics for data analysis is still rare, as
  - the focus of statistical education is often not on data analysis.
  - it is not taught at universities (except for some misfits).
  - there is not much continuity in tools yet.
  - it still lacks possible formalization.



Conclusion

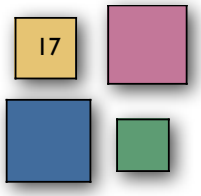
- Graphics for data analysis is still rare, as
  - the focus of statistical education is often not on data analysis.
  - it is not taught at universities (except for some misfits).
  - there is not much continuity in tools yet.
  - it still lacks possible formalization.

- Formalization can be done for some parts of the techniques involved (IMHO it does not make sense to go for a global model).



Conclusion

- Graphics for data analysis is still rare, as
  - the focus of statistical education is often not on data analysis.
  - it is not taught at universities (except for some misfits).
  - there is not much continuity in tools yet.
  - it still lacks possible formalization.
- Formalization can be done for some parts of the techniques involved (IMHO it does not make sense to go for a global model).
- Unfortunately statistical computing degraded to implementing yet another package for R ⇒ not much hope for new software.



Conclusion

- Graphics for data analysis is still rare, as
  - the focus of statistical education is often not on data analysis.
  - it is not taught at universities (except for some misfits).
  - there is not much continuity in tools yet.
  - it still lacks possible formalization.
- Formalization can be done for some parts of the techniques involved (IMHO it does not make sense to go for a global model).
- Unfortunately statistical computing degraded to implementing yet another package for R ⇒ not much hope for new software.
- Apart from a formal description in a mathematical sense, the way of teaching concepts and tools via case studies seems to be quite promising …

"Commercial Break"