Statistical Computing / Computational Statistics

Friedrich Leisch
Martin Theus

In Short

Statistical Computing
/
Computational Statistics

~~S~~ ~~C~~ / ~~C~~ ~~S~~

1 / 1

Portugal : Nederlands

Definition (for now)

- **Statistical Computing**

  Providing the computational tools for statistics by using tools and techniques from computer science

- **Computational Statistics**

  Doing statistics by using computer based tools

Statistics & Computing: Past

- Started as “Offline – Computing”
- Interaction was often limited to the operator …
- Stat Computing = Software Development
- Stat Graphics
  - a step back for a long time
  - then BLNFYs
  - most system still inherit their graphics system from pen plotter times
- Many heterogeneous “libraries”
- Stat Computing = Computational Statistics

What gained Statistics from Computing?

- What can be done without a computer?
  - Tests
  - Models
  - Graphics
- What can be done better with a computer?
  - Tests, Models, Graphics
  - Tests, Models, Graphics for larger data
  - …
- What can not be done without a computer?
  - Advanced Models
  - Simulations
  - Interactive Graphics and Computing
  - Statistical learning algorithms

(Mathematical) Statistics: Two Views

- An application of mathematics
  - distribution based
  - small to midsize samples
  - emphasis on significance

  - “fitting data to models”
  - computing helps to turn models into software

- Supporting Data Analysis
  - data based
  - complex data rather than samples
  - emphasis should be on relevance

  - “fitting models to data”
  - computing helps to generate hypotheses

Stat Learning & Computer Science

- Why computer scientists steal the show from statisticians
  - better marketing
  - database savvy
  - ...

- What is Data Mining?
  - Pregibon: “Statistics at scale and speed”
  - Ripley: “Machine learning is statistics minus any checking of assumptions”
    “Data Mining is machine learning with better marketing”

- Is Stat Computing shifting away from Statistics?

Statistics & Computing: Today

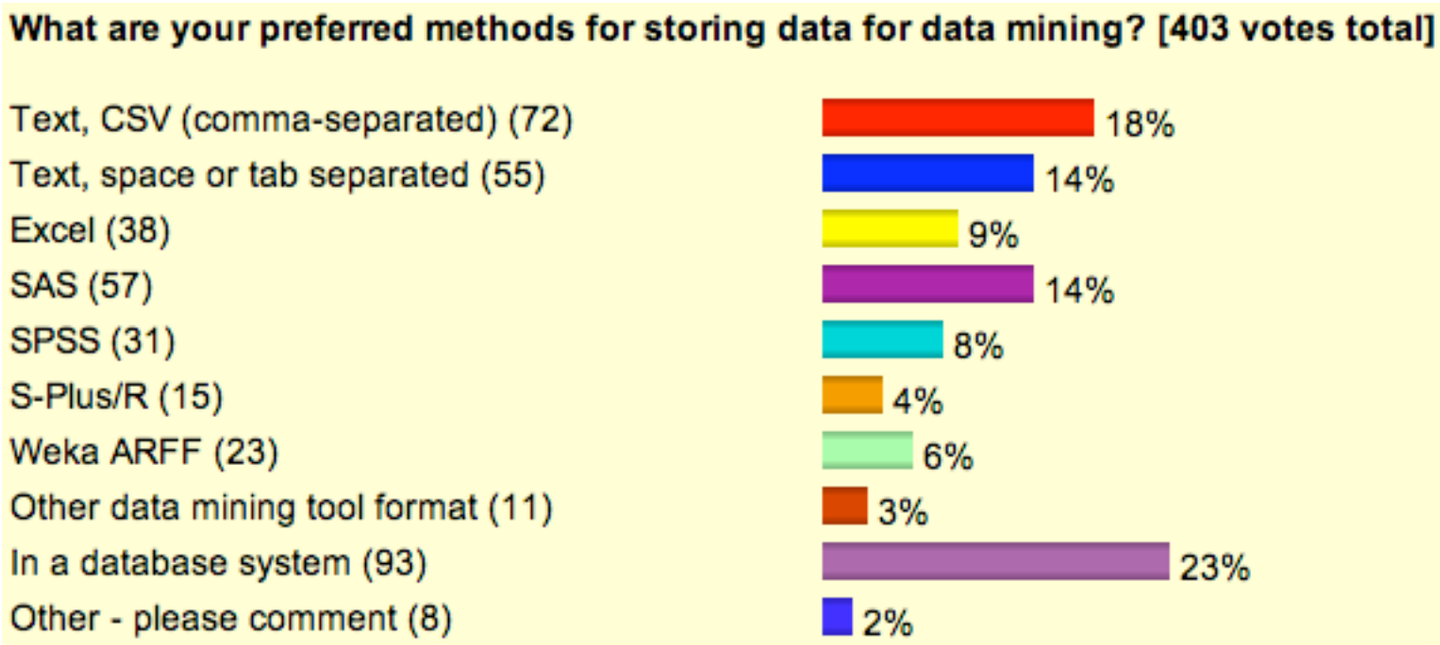
- "Everybody does it" with/for R (true only for stat computing)
- Computational Statistics still done with SAS, SPSS, … ?
- Fewer and fewer (standalone) software projects in statistical research.
  Is this good or bad?
- Computational Statistics → 100%
- Statistical Computing → ϵ, ϵ small; how small?
- Impact of research on commercial software is still rather small.

Who does Statistical Computing?

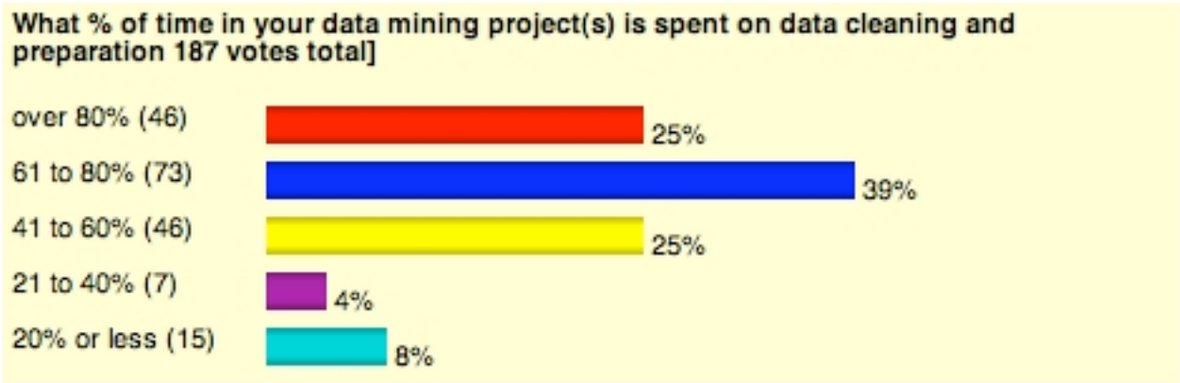
- Past: “real programmers”
- Today:
  - Statisticians
  - Domain Experts
  - Computer Scientists
  - more and more non-programmers
- Future:
  - Is writing an R function statistical computing?
  - Is writing an R package statistical computing?
  - Is writing an R successor statistical computing?

Software for Data Mining: Web Survey

- Data Sources



- Data Preparation



Software for Data Mining: Web Survey (cont.)

- Techniques

Data mining/analytic techniques you use frequently: [784 votes total]

Decision Trees/Rules (107)	14%
Clustering (101)	13%
Regression (90)	11%
Statistics (80)	10%
Visualization (63)	8%
Neural Nets (61)	8%
Association rules (54)	7%
Nearest Neighbor (34)	4%
SVM (Support vector machine) (31)	4%
Bayesian (30)	4%
Sequence/Time series analysis (26)	3%
Boosting (25)	3%
Hybrid methods (23)	3%
Bagging (20)	3%
Genetic algorithms (19)	2%
Other (20)	3%

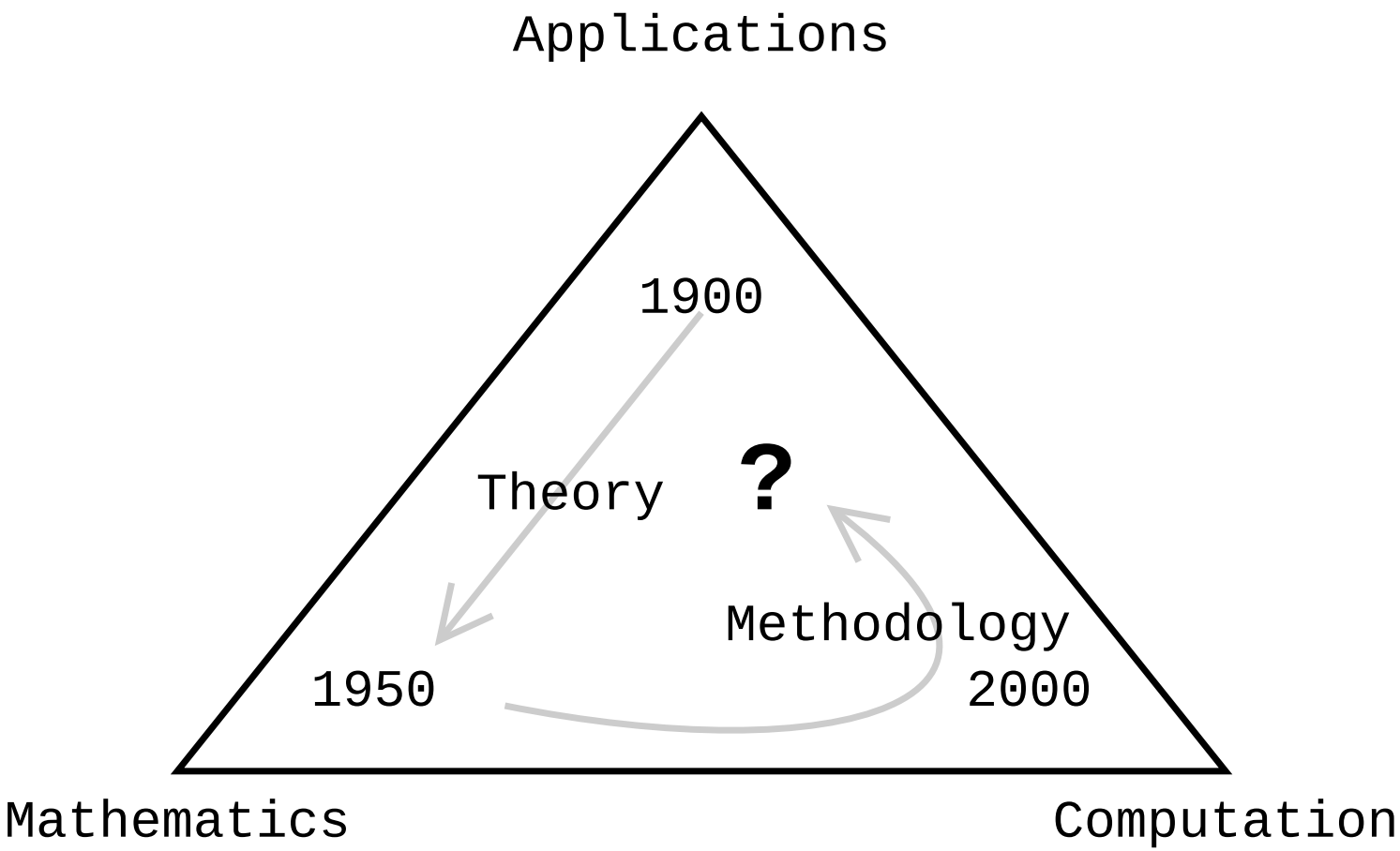
- Tools

Data mining/analytic tools you used in 2006: [561 voters]

CART/MARS/TreeNet/RF	159 (72 alone)
SPSS Clementine	127 (47 alone, 46 with SPSS)
SPSS	100 (5 alone, 46 with Clementine)
Excel	100 (3 alone)
KXEN	90 (75 alone)
your own code	77 (1 alone)
SAS	72 (3 alone, 13 with E-Miner)
Weka	62 (7 alone)
R	53 (5 alone)
MATLAB	41 (5 alone)
other free tools	39 (3 alone)
SAS E-Miner	37 (9 alone, 13 with SAS)
SQL Server	32 (3 alone)
other commercial tools	31 (7 alone)
Oracle Data Mining	20 (13 alone)
Insightful Miner/ S-Plus	20 (0 alone)
C4.5/C5.0/See5	18 (0 alone)
Megaputer	16 (14 alone)
Statsoft Statistica	13 (2 alone)
Angoss	8 (2 alone)
Mineset (PurpleInsight)	5 (3 alone)
Eudaptics Viscovery	5 (3 alone)
Xelopes	4 (4 alone)

Statistics and Computers

Efron (2001): Statistics in the 20th century



Persistent Communication

Persistent Communication

Persistent Communication

Persistent Communication

Persistent Communication

What is statistical software?

- A program, which takes numbers as input and creates tables (and figures)?
- A (collection of) program(s) for exploration, inference and modelling?
- A tool for administration, manipulation and analysis of data?
- A medium of communication with CPU (graphics card, printer, . . .) of our computer?

Communication forms

Pictoral languages: Easy to learn, often universally understood, limited in complexity and expressions.
Examples: Austrian road signs, Apple-GUI, etc.

Written languages: Hard to learn, need choice of language, almost unlimited in complexity and expressions.
Examples: Austrian German, UNIX-Shell, etc.

Iliad and Odyssey by Homer have survived until today. The mythology of ancient Egypt may have been as complex as the one of ancient Greece, but try the exercise of encoding the first chapter of the Iliad in icons without loss of information!

OK, writing it as a shell script isn't that much easier . . .

. . . unless the script is well documented!

Relevance for data analysis

- Statistical software is a tool to tell the computer how to analyse your data.

- For simple analyses simple forms of communication are sufficient, but even the most complex GUI has only a finite number of submenus.

- Only programmable environments are (almost) unlimited.

- S has won the oscar for software systems (ACM award), S·S and S··S are still waiting for it.

Relevance for data analysis

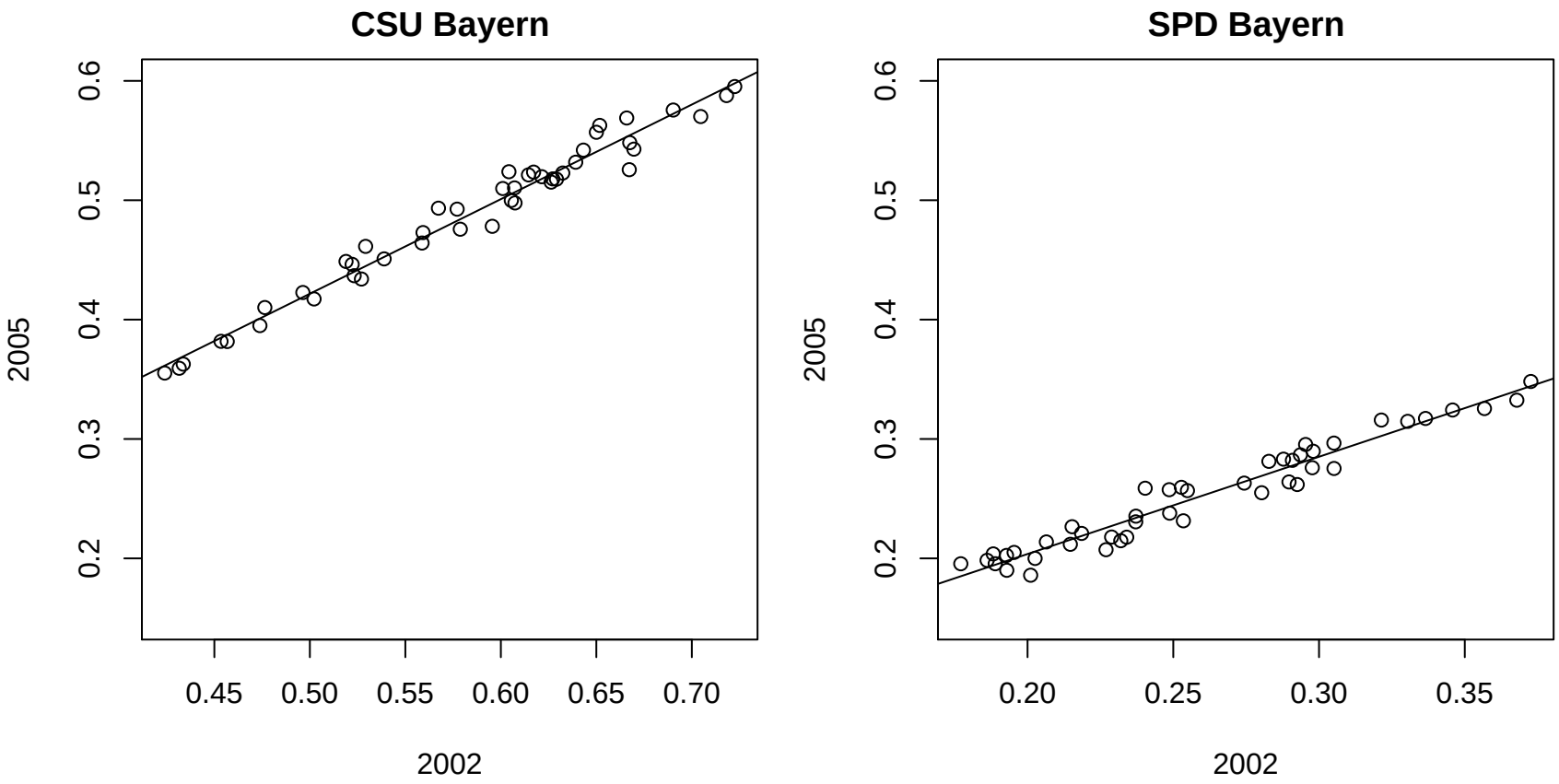
- R has become the probably most popular tool for programming with data within academic statistics over the last years.

- Using a common language (Latin, English) has always been a key to success for science to avoid reinventing the wheel.

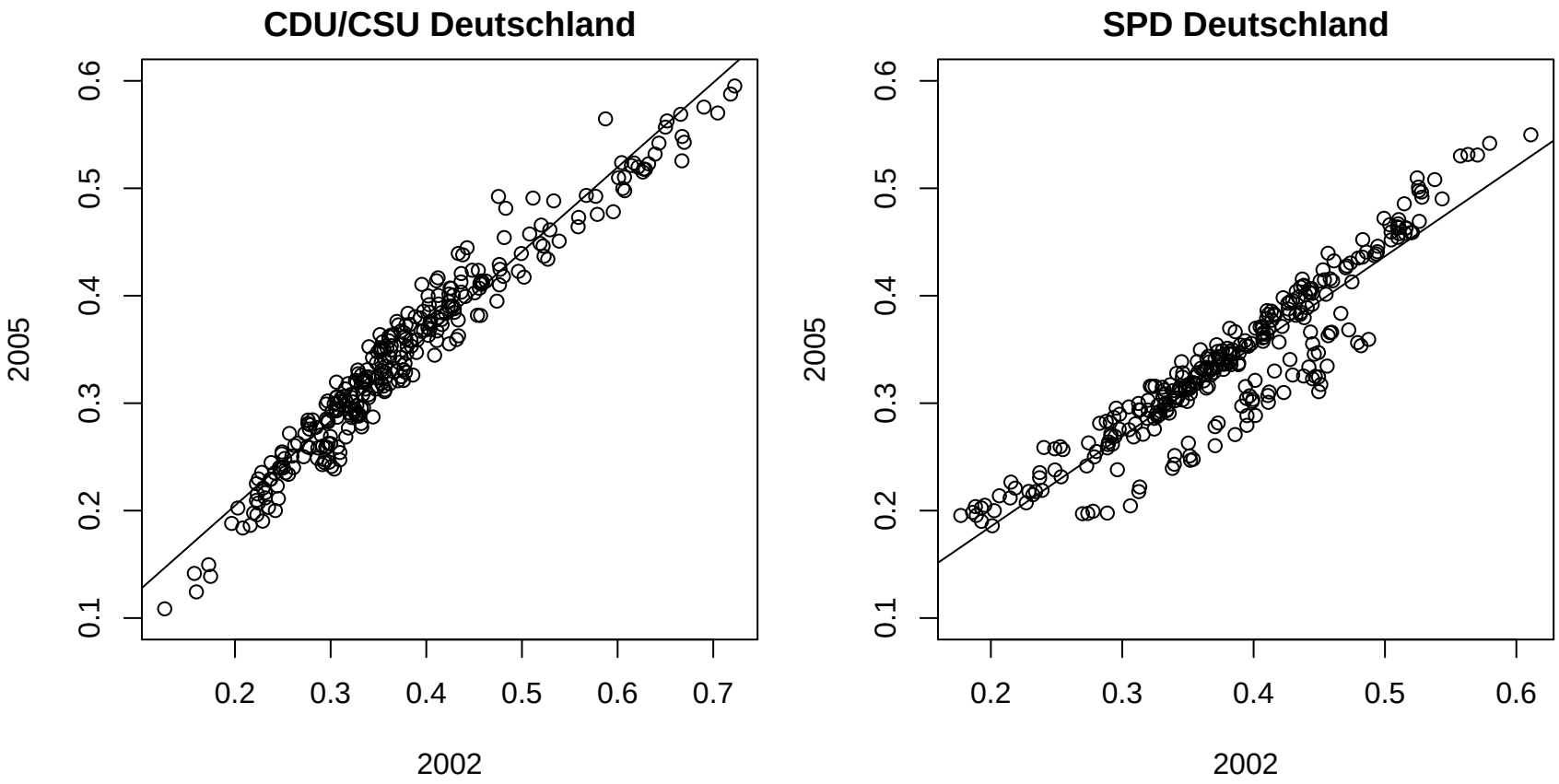
- This is the first time that we (as a community and as a profession) own the computational environment we use for more and more of our data analyses.

- Publishing data and R code makes statistical research truly reproducible.

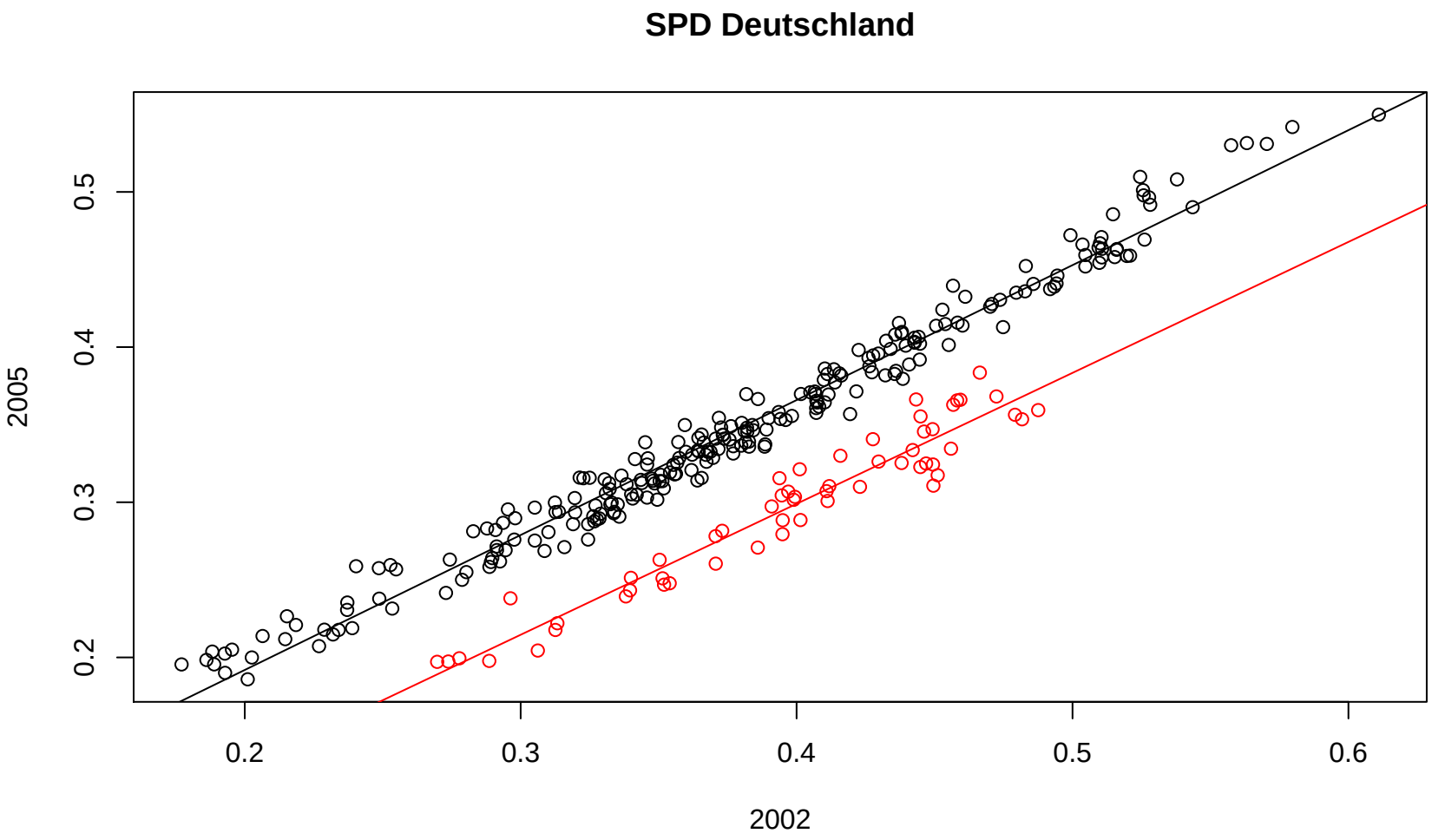
Ex: German elections 2005



Ex: German elections 2005



Ex: Mixture model for SPD



Ex: Mixture model for SPD

```
> table(btw2005$Bundesland, cluster(f1))
                          1  2
  Baden-Württemberg       0 37
  Bayern                  0 45
  Berlin                  0 12
  Brandenburg            10  0
  Bremen                  0  2
  Hamburg                 0  6
  Hessen                  0 21
  Mecklenburg-Vorpommern  7  0
  Niedersachsen           0 29
  Nordrhein-Westfalen     0 64
  Rheinland-Pfalz         0 15
  Saarland                4  0
  Sachsen                16  1
  Sachsen-Anhalt         10  0
  Schleswig-Holstein      0 11
  Thüringen               9  0
```

R only???

OK, until now I have done the expected to create a (hopefully) strong case advocating the usage of R.

Who would have problems turning the following (spoken) directions into proper R code?

```
zett ist gleich l m klammer auf ypsilon tilde x beistrich
  data ist gleich d klammer zu
```

Several of my students in Munich had and it took me quite some time to figure out why.

Want some sweets?

- I guess most people will agree that the existening differences of laguages, cultures, etc. (E-DOLCE) makes the world a much more interesting place to live in.

- E-DOLCE wastes resources, but is also a source of inspiration.

- (A dash of) Breiman: Machine learning has reinvented a lot of things that statisticians already knew, but it also provided many new ideas which have subsequently been adopted by us.

- There are many good reasons, why many road signs feature icons rather than words.

Challenges for the future

- It took mankind quite some time to develop the languages we speak and write today, it would be naive to assume that we already have the definite language for statistical data analysis.

- We need better (and most likely radically different) ways to
  - specify large and complex models
  - deal withy large data sets
  - make our tools safely available to practitioners
  - YOUR INPUT

Statistical Graphics: Two Worlds?

- Stat graphics is widely used; but mainly for diagnostics
- Most diagnostic plots do not really visualize the model, but residuals or other model characteristics
- Understanding diagnostic plots is usually more complex than understanding plots of the raw data
- Chart Junk!
  There are as many stupid graphics out there as stupid models,
  With graphics one would hope to find out more easily …
  (e.g. check the graphics section of “The American Statistician” regularly)
- The data visualization community often show us how not to do it

Graphic for Data Cleaning

- Data cleaning and preparation takes most of the time for most of the projects
- Hardly taught in any statistical curricula
- Finding “the round tag in the square hole” is far easier with graphics than with any other tool
- Undiscovered error and artifacts in the data often end up as “the signal” in the data, or at least will obscure the real thing
- What are standard tools and procedures for data cleaning? (or do we mostly rely on common sense here?)

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 not easy to teach
- a case study based approach is usually most successful

- 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 (at least some) software that helps

Interactive Computing in Statistics

- Command line input with (almost) instantaneous response was a big advance
- Interaction with command line is limited
- Using graphics, new ideas are usually faster generated than you can type the next error-free command
- Direct manipulation of statistical objects (be it data, models or graphics) speed things up drastically, but that has its limits, too
- Multiple views of statistical objects are crucial
  (although we might work sequentially, having results in parallel will better support cognitive reasoning)

On the Interface

- Graphical user interfaces (GUIs) started the “Desktop Revolution”
- As with any revolution, things fail, if the initial impetus is lost …
  (There **are** GUI guidelines and books from D.A. Norman and J. Nielsen are fun)
- There was a great momentum in stat computing 20 years ago to leverage this technological advance
- Most software projects either died or stalled
  (DataDesk is still great, but essentially the same software since 1994)
- What about R’s user interface?

Too expensive & and not rewarded
(or do you know someone who got tenure for great software?)

What makes Graphics interactive?

- **Direct manipulation is at the core**

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

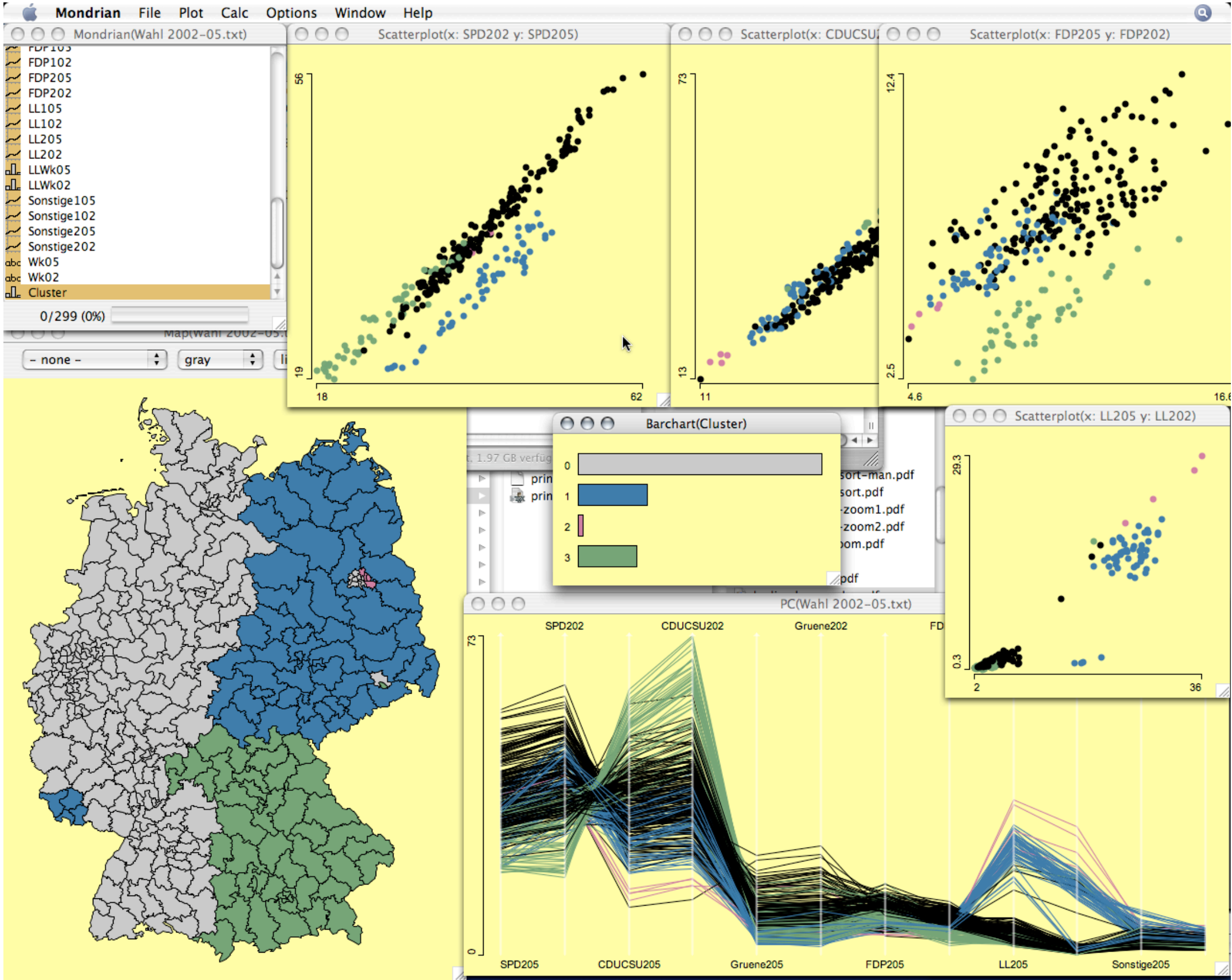
  – Selection of subgroups, i.e., conditioning
  – Modification of plot parameters

- **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**

- **Linking**

  Attributes can be linked across multiple views
  (graphics, summaries, models, …)

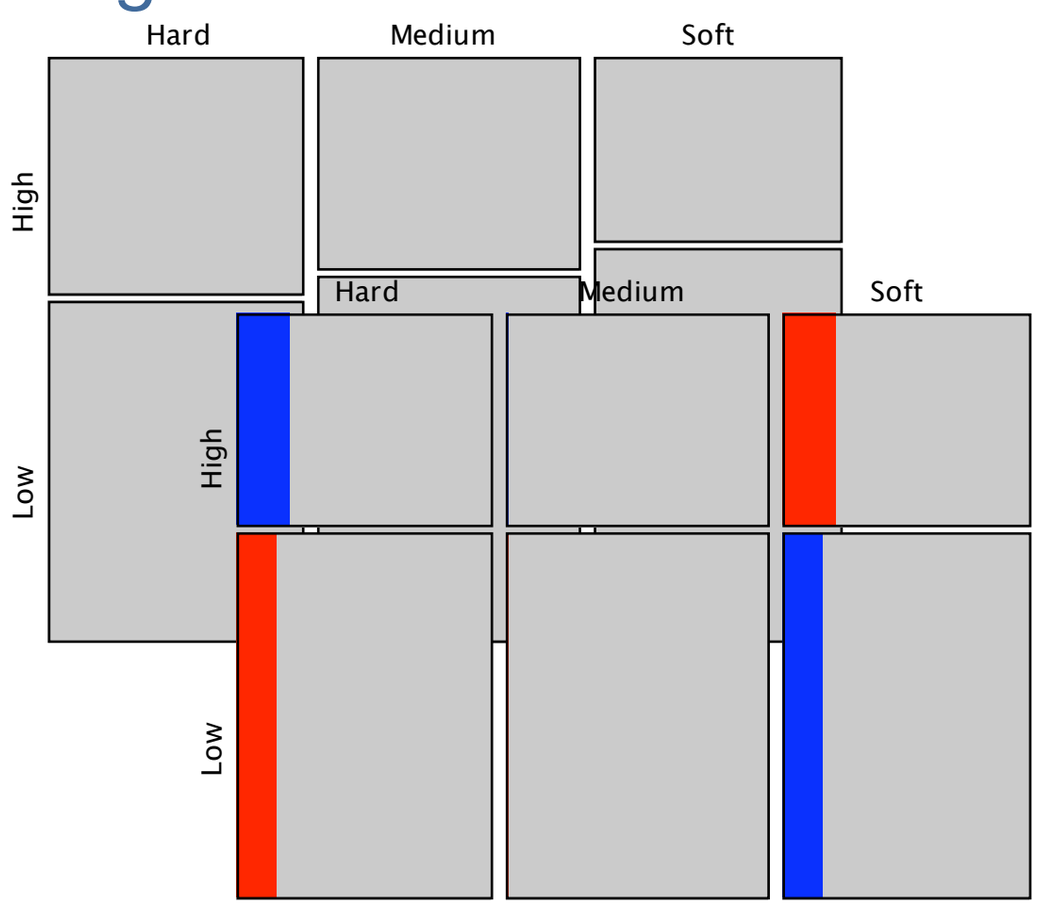
DEMO


Mondrian File Plot Calc Options Window Help
Mondrian(Wahl 2002-05.txt)
FDP102
FDP205
FDP202
LL105
LL102
LL205
LL202
LLWk05
LLWk02
Sonstige105
Sonstige102
Sonstige205
Sonstige202
Wk05
Wk02
Cluster
0/299 (0%)
- none -
gray
Scatterplot(x: SPD202 y: SPD205)
Scatterplot(x: FDP205 y: FDP202)
Barchart(Cluster)
Scatterplot(x: LL205 y: LL202)
PC(Wahl 2002-05.txt)
SPD202
CDUCSU202
Gruene202
SPD205
CDUCSU205
Gruene205
FDP205
LL205
Sonstige205

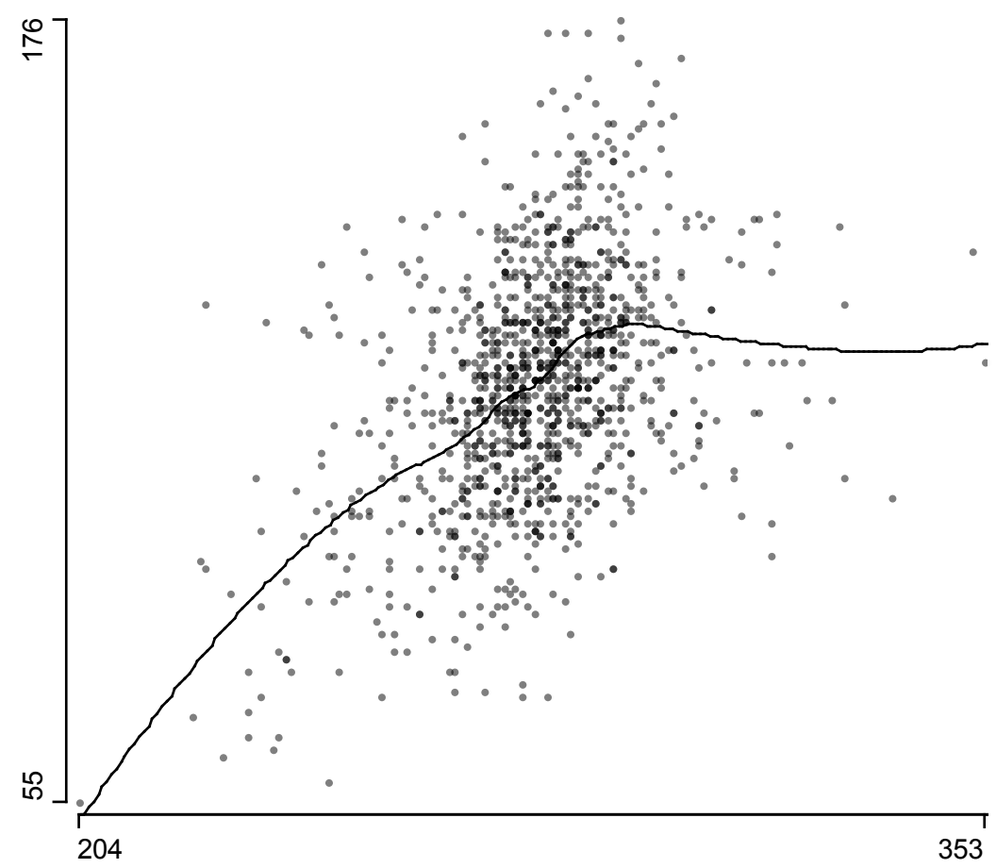
Enhancing Graphics

- Add results from numerical methods from classical statistics to graphical methods (or vice versa).
- Idea: Make visible what is not immediately visible or quantifiable.
- Examples

log-linear models



Scatterplot smoothers



Main Questions ...

- Statistical Computing and Computational Statistics started as one; what is the state today?
- What is the relation of computations and graphics in statistics; are they both computing?
- Does statistical computing leverage current technologies from computer science?
- Is the gap between users and developers increasing or decreasing?
- Which side are you on; or where do you want to be next week?