1001 Graphics

Martin Theus

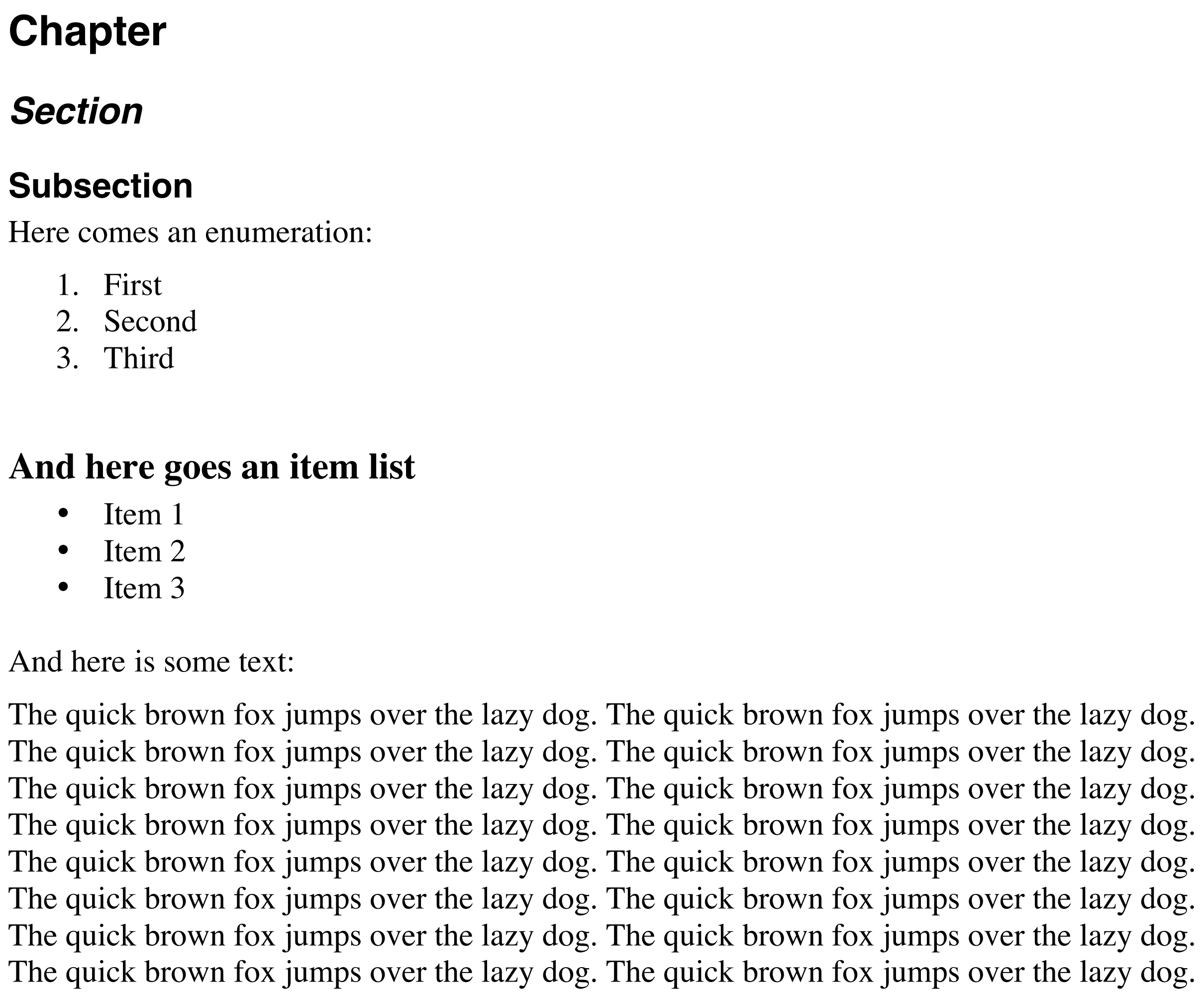
martin.theus@math.uni-augsburg.de

Why talking about "Defaults"?

- Living in a world with free global markets, we have to face choices
- Problems with choices:
  - the number of options is often far greater than the non-expert can handle.
  - unlikely that we always want to take care about all settings
- Consequence ⇒ we have to use Defaults!
- Examples:
  - Typography (MS Word vs. LaTeX vs. Quark XPress vs. InDesign …)
  - Statistical methods and statistical graphics and their parameters
  - …
- In principle two situations
  - There does not exist a global optimum ⇒ choose as you like
  - There is a recognized "best solution" ⇒ take the default and you are far off!

Example: Typography

- Word



Chapter

Section

Subsection

Here comes an enumeration:

1. First
2. Second
3. Third

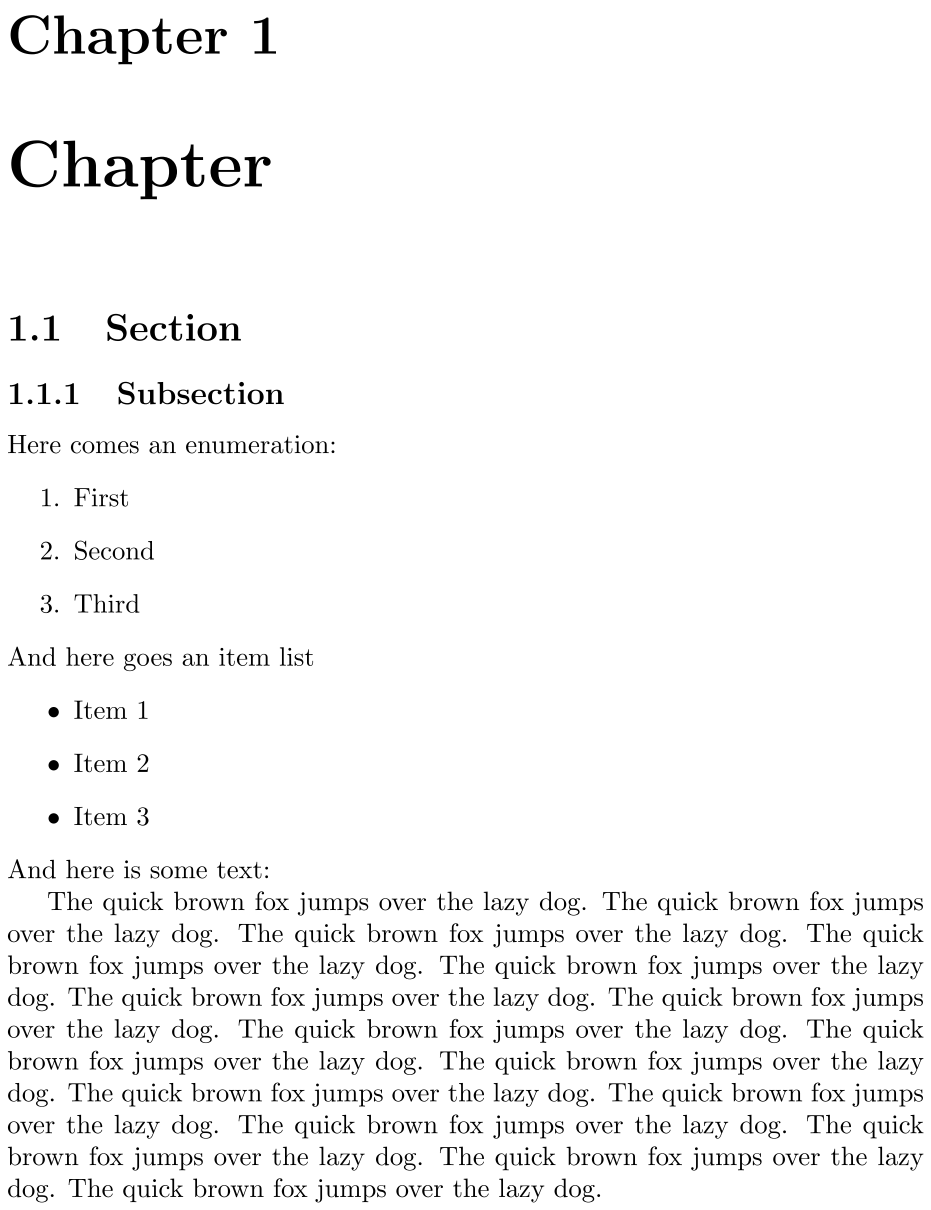
And here goes an item list

- Item 1
- Item 2
- Item 3

And here is some text:

The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.

- LaTeX



Chapter 1

Chapter

1.1 Section

1.1.1 Subsection

Here comes an enumeration:

1. First
2. Second
3. Third

And here goes an item list

- Item 1
- Item 2
- Item 3

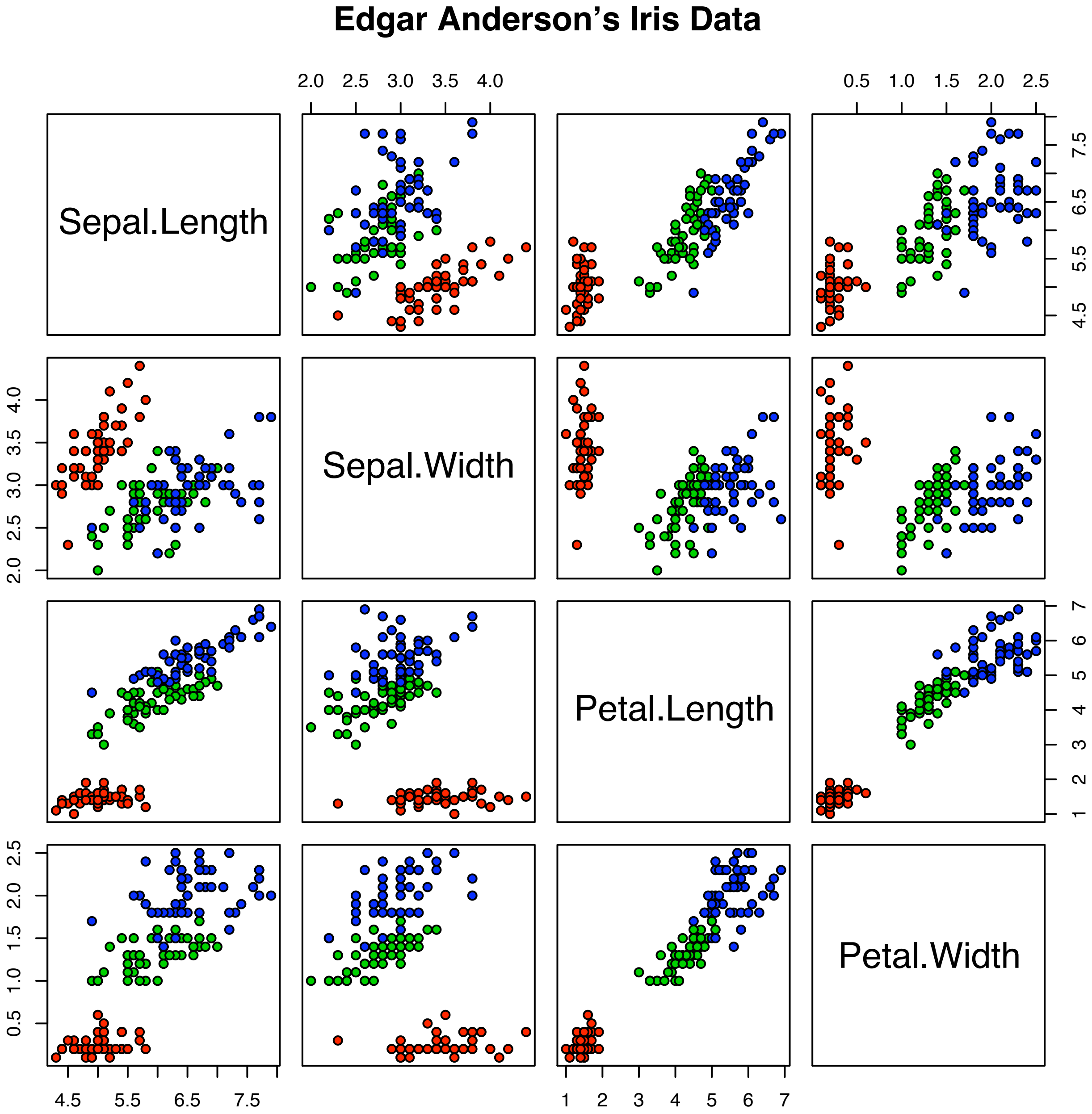
And here is some text:

The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.

Defaults in Scatterplots

- Looks good at first sight

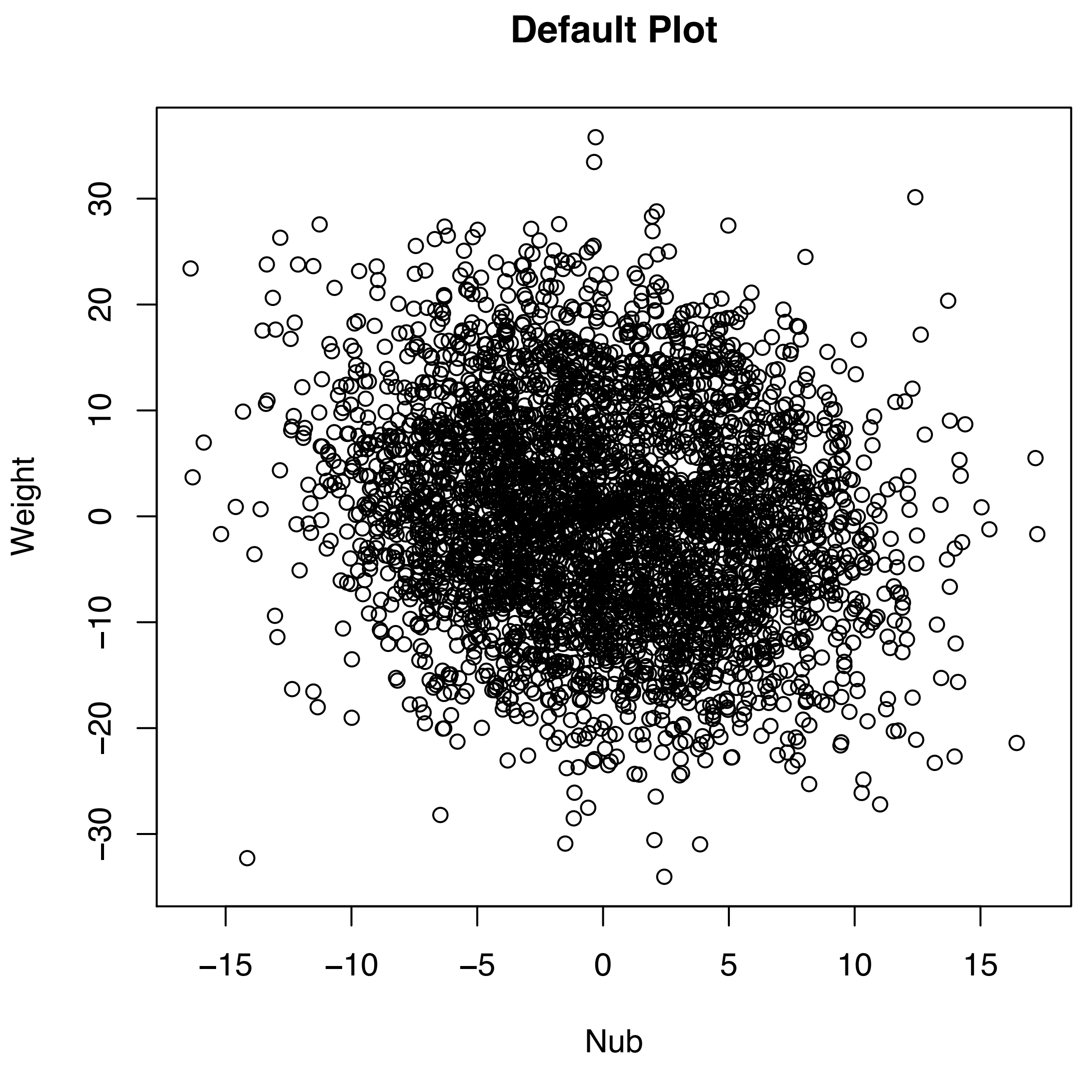
```
pairs(iris[1:4],
      main = "Edgar Anderson's Iris Data",
      pch = 21,
      bg = c("red",
             "green3",
             "blue")[unclass(iris$Species)])
```



What if Anderson’s Lab would have had the money to look at 10 species 400 samples each?

Pollen Data

- From "ASA Data Competition 198?"

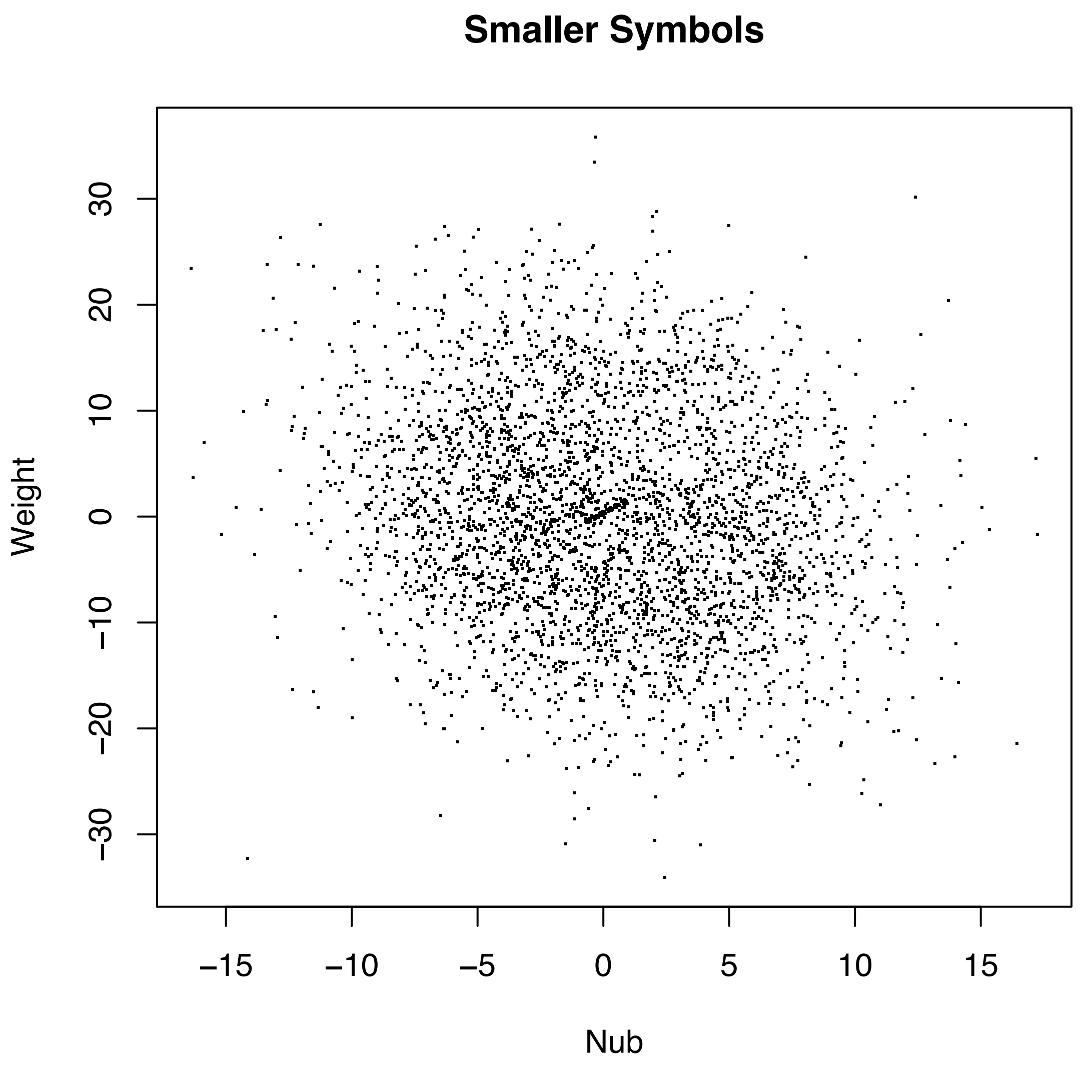


3,848 simulated cases in 5 dimensions

What do you see?

Pollen Data

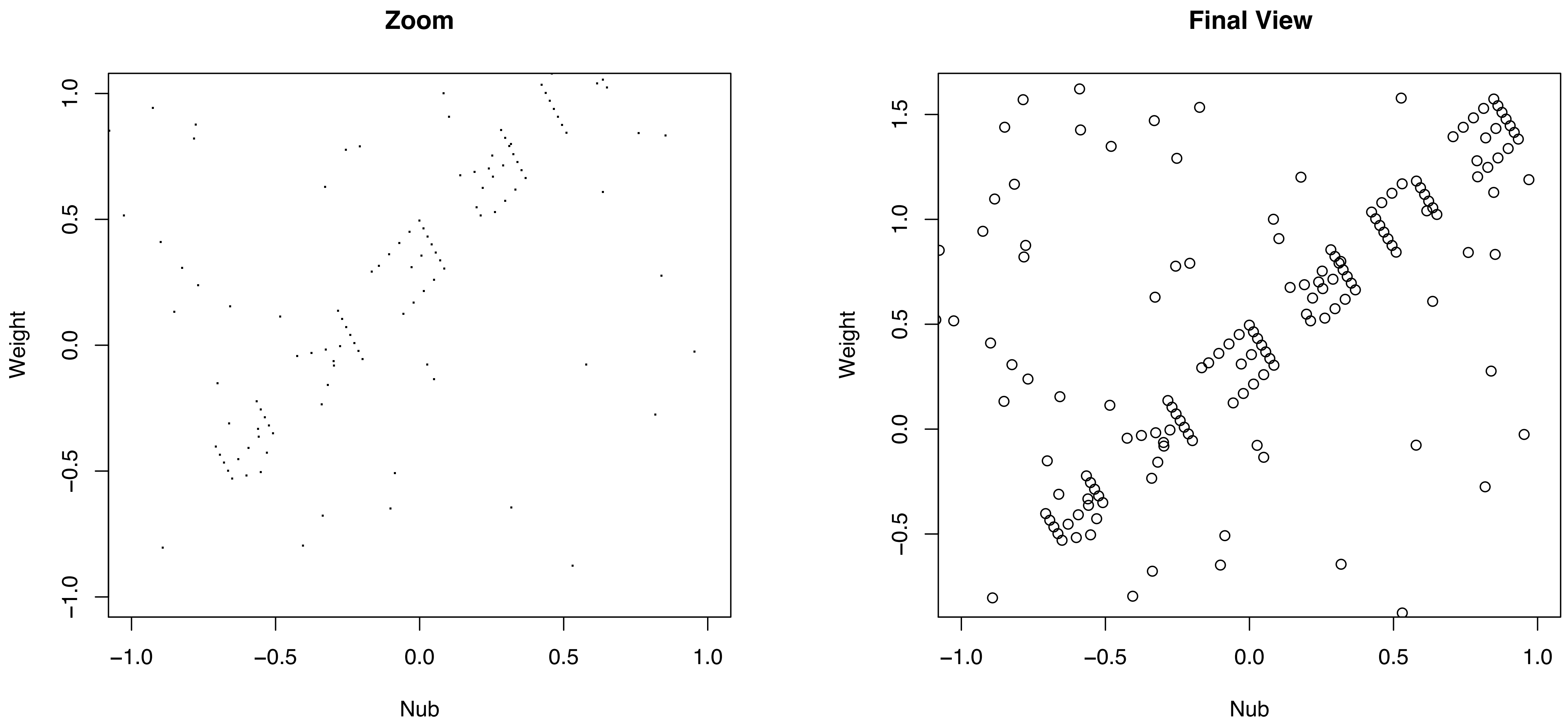
- Make the plot symbol smaller: Aha!



Something's strange in the center

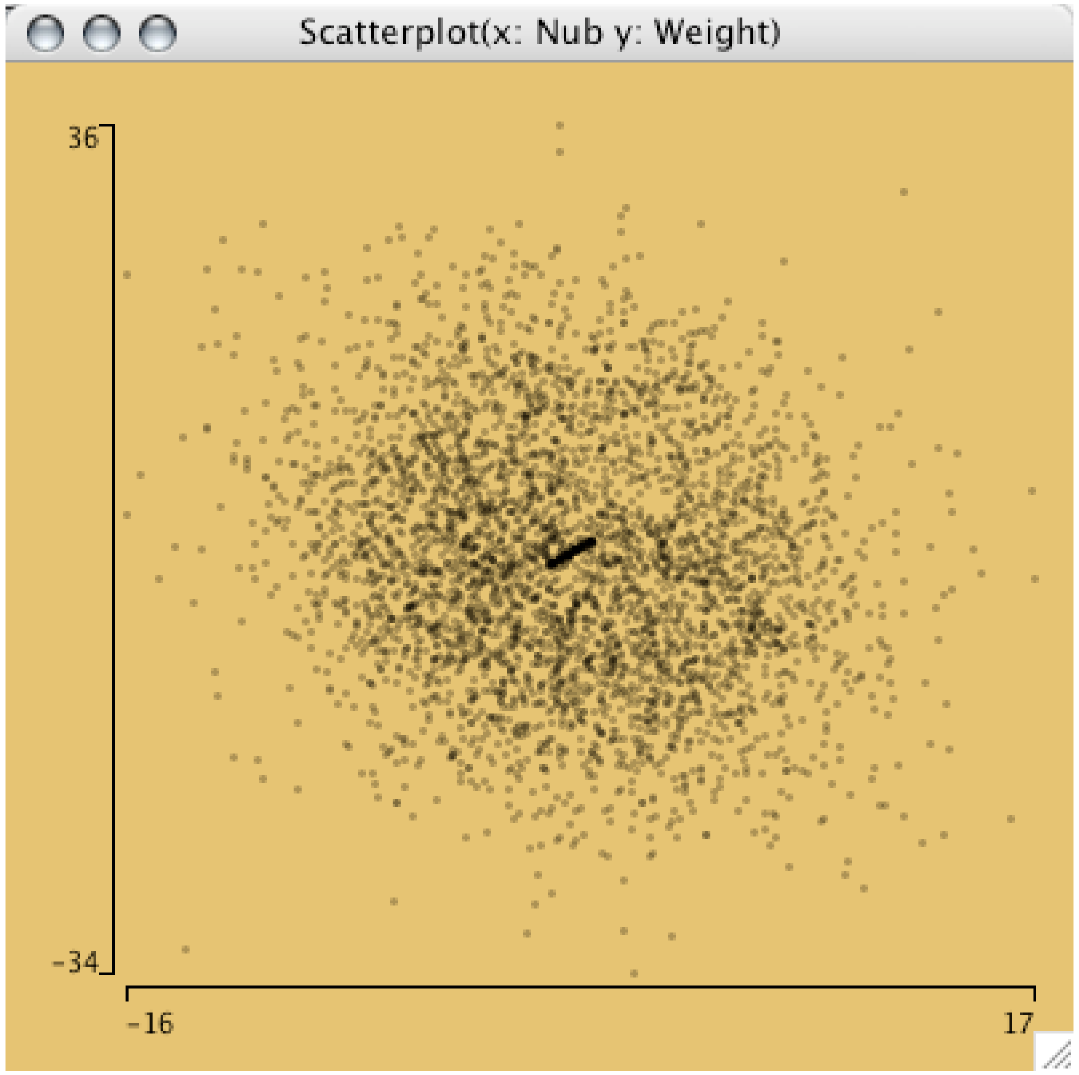
Pollen Data

- Zoom in and Fine Tune



Pollen Data: Summary

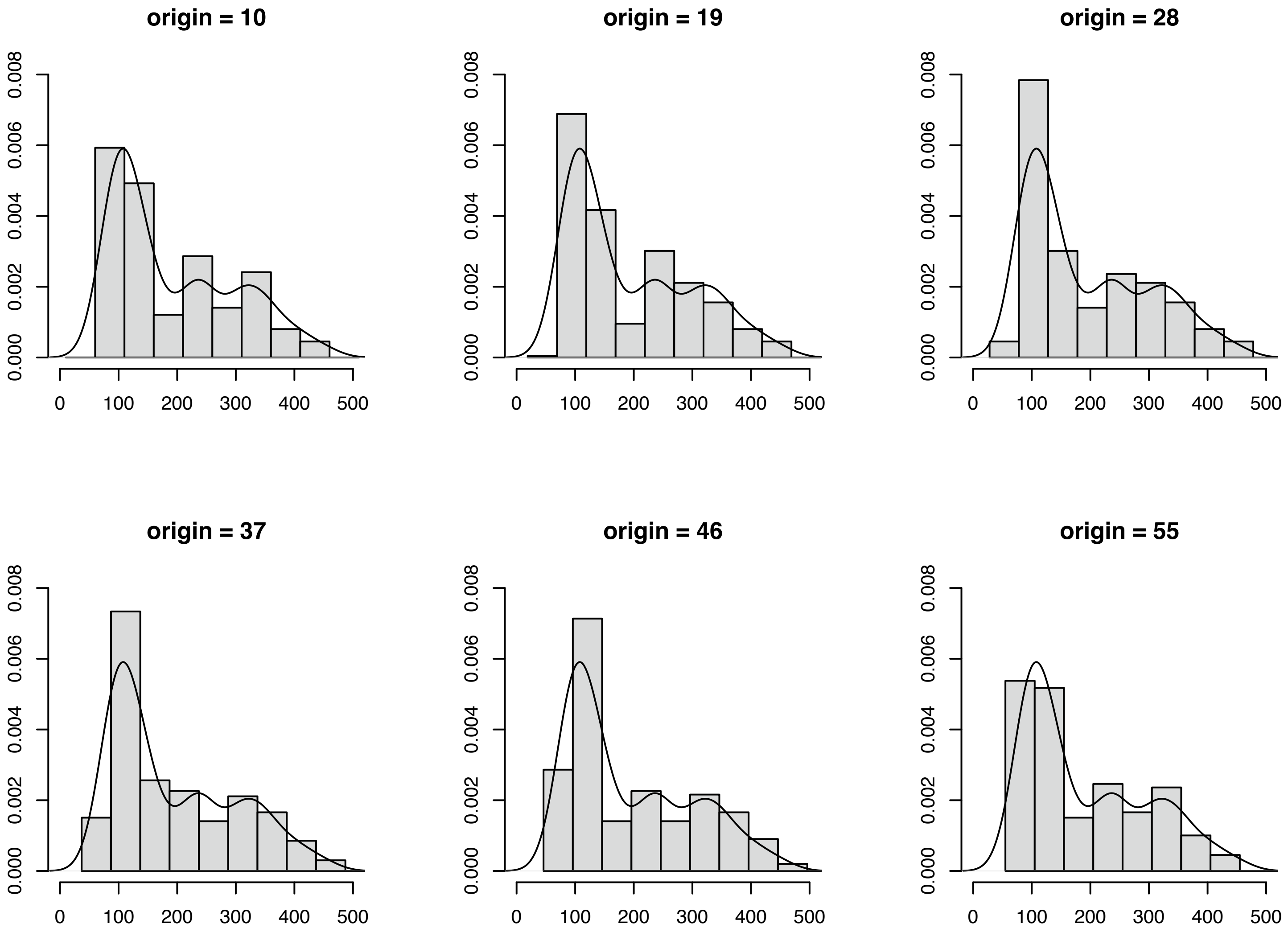
- R default plot symbol works for less then 100 points … ☹
- Symbol size should adjust to the size of the data set …
- … actually the size must be adjusted to the #points plotted



- Easy solution: Plotting with α-transparency

The Histogram

- There are still people around, who think that histograms are good at visualizing densities resp. distributions of real data.
- 6 histograms of displacement of ~400 cars with different origins

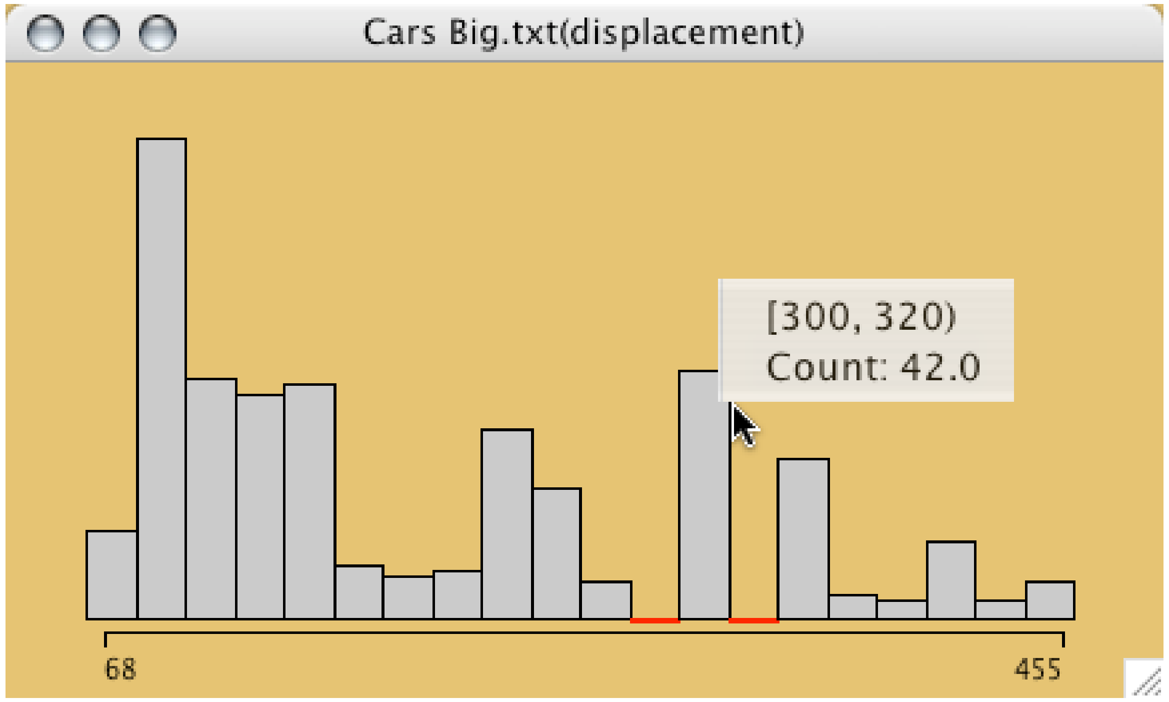


Cars Data: Displacement

- Displacement is sampled from at least 3 processes:
  4, 6 and 8 cylinder cars,
  and has accumulation points:
  1.6l, 1.8l, 2.0l, …

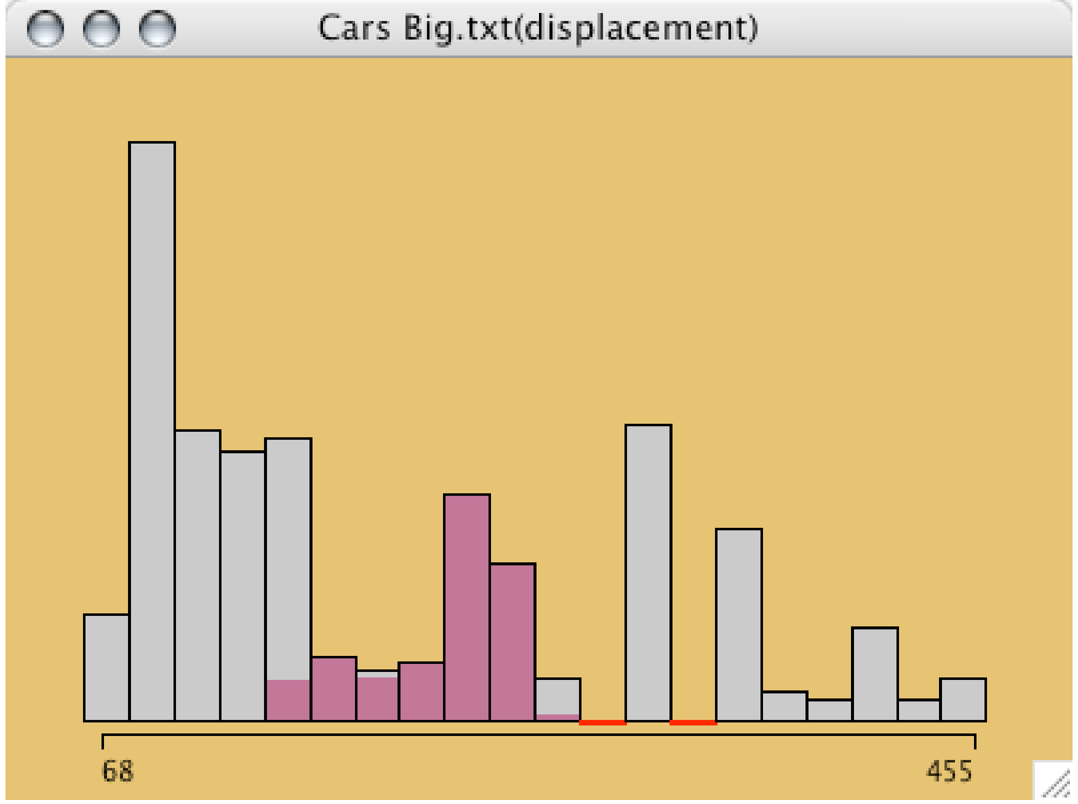
- ⇒ interactively look for a suitable visualization …

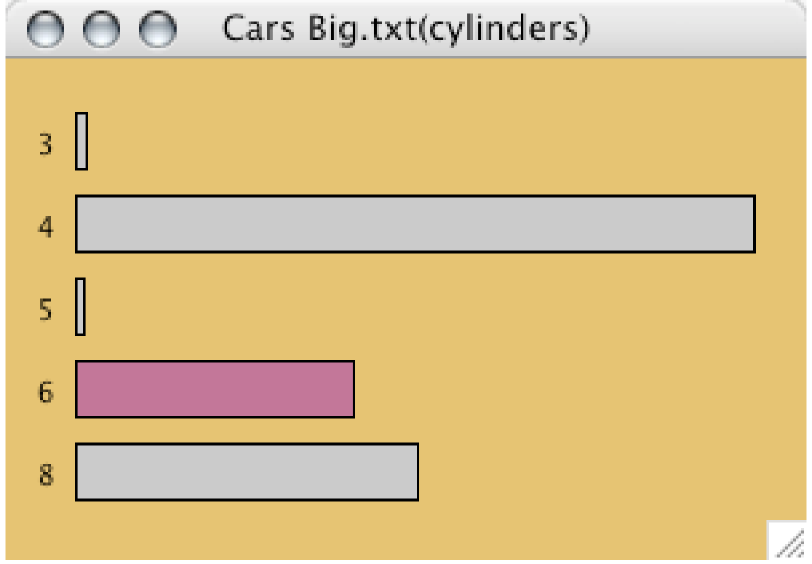
forget the density



Cars Data: Interpretation

- Linking Displacement and Cylinders





- ⇒ histograms are good for selection and highlighting

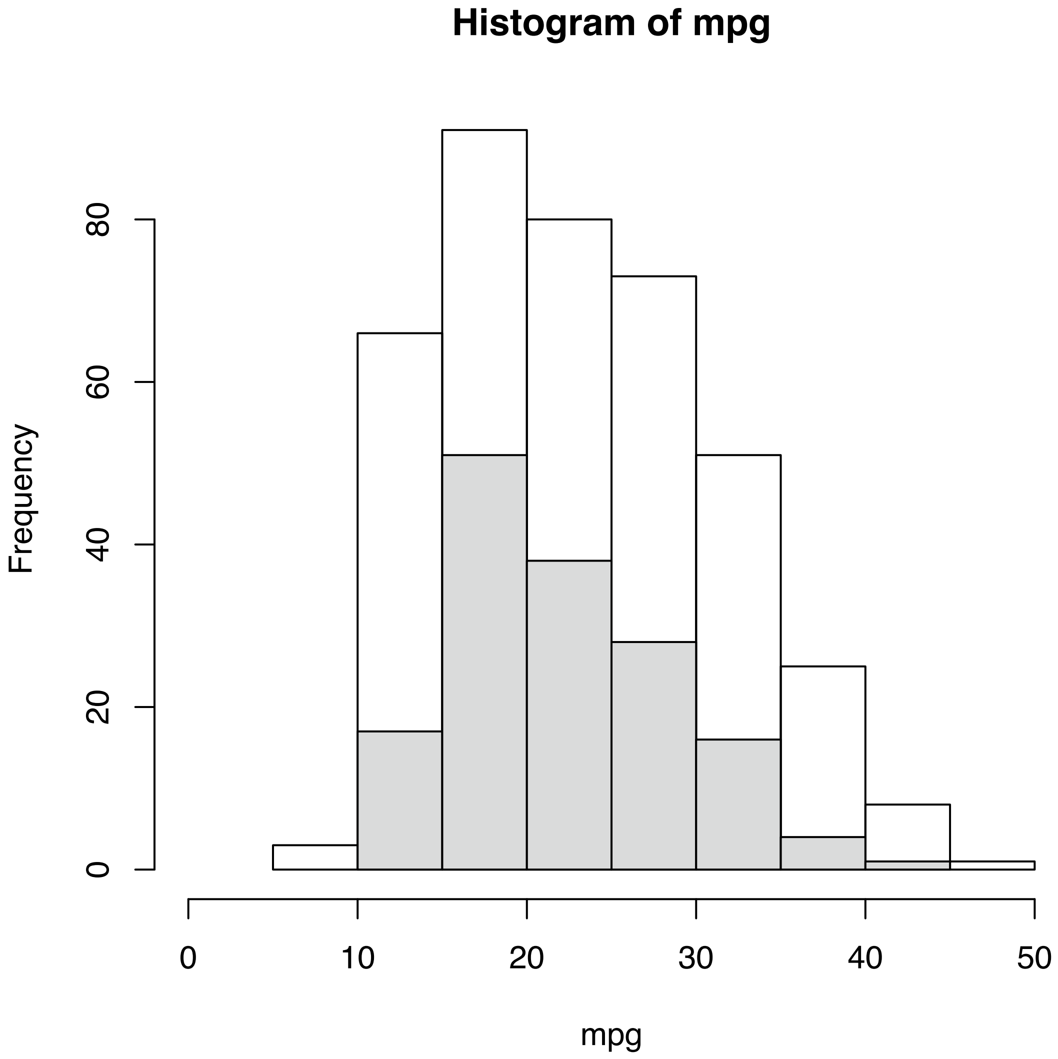
Histograms: Summary

- With real data there is hardly any optimal anchor point and bin width for histograms
- If we really look for a density, density estimators are always the better choice
- The best default for anchor and bin width is probably "no default"
- Interactivity can help
- Histograms are especially useful for linked-highlighting

Detour: Spinograms

- Selection and highlighting in histograms can be misleading:

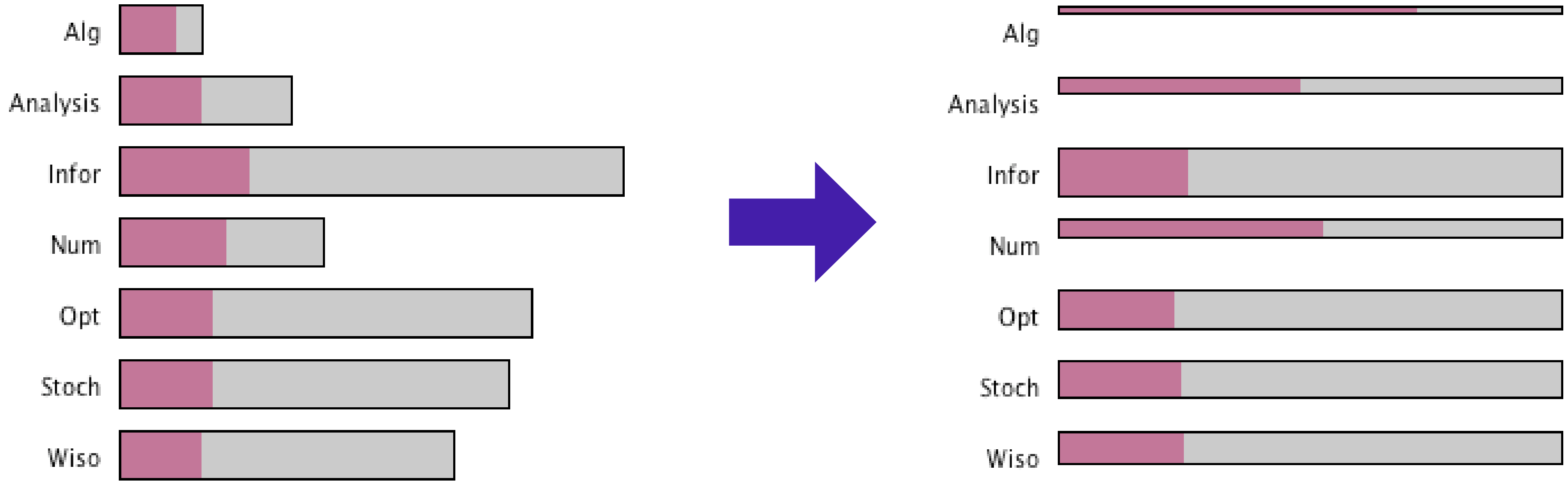
  What is the distribution of 'mpg' of model makes '74-'78?



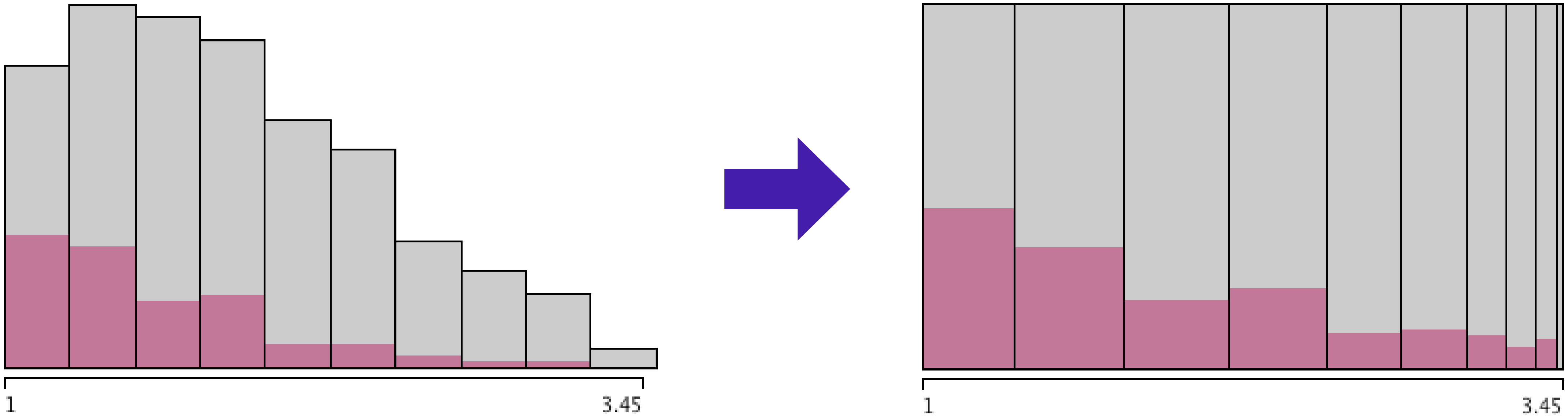
You can't really tell,
but spinograms can help!

Spinograms: Definition

- Let’s first look at spineplots:

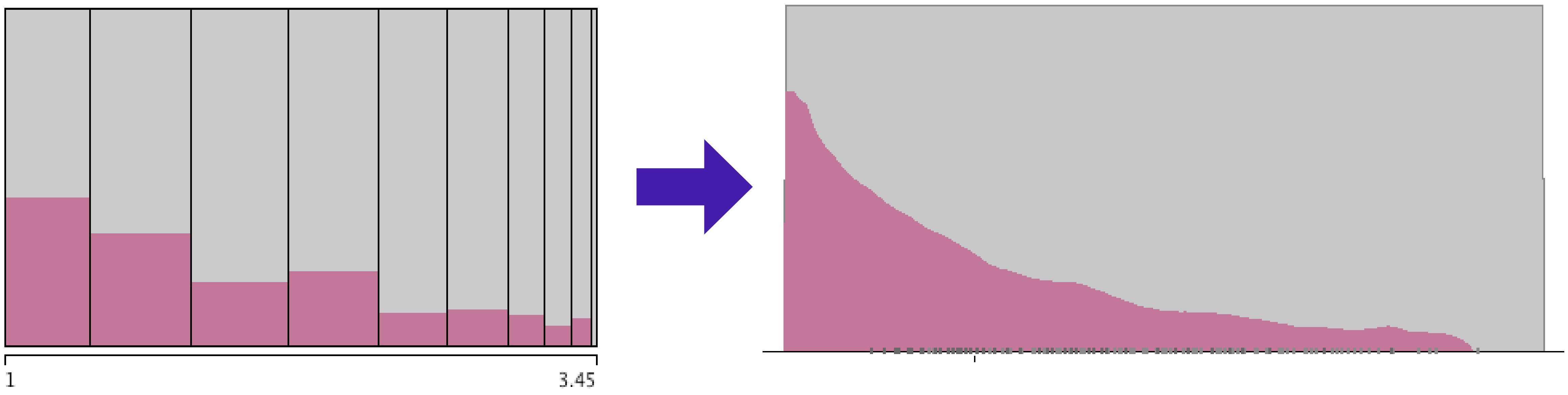


- Let’s make the same with histograms



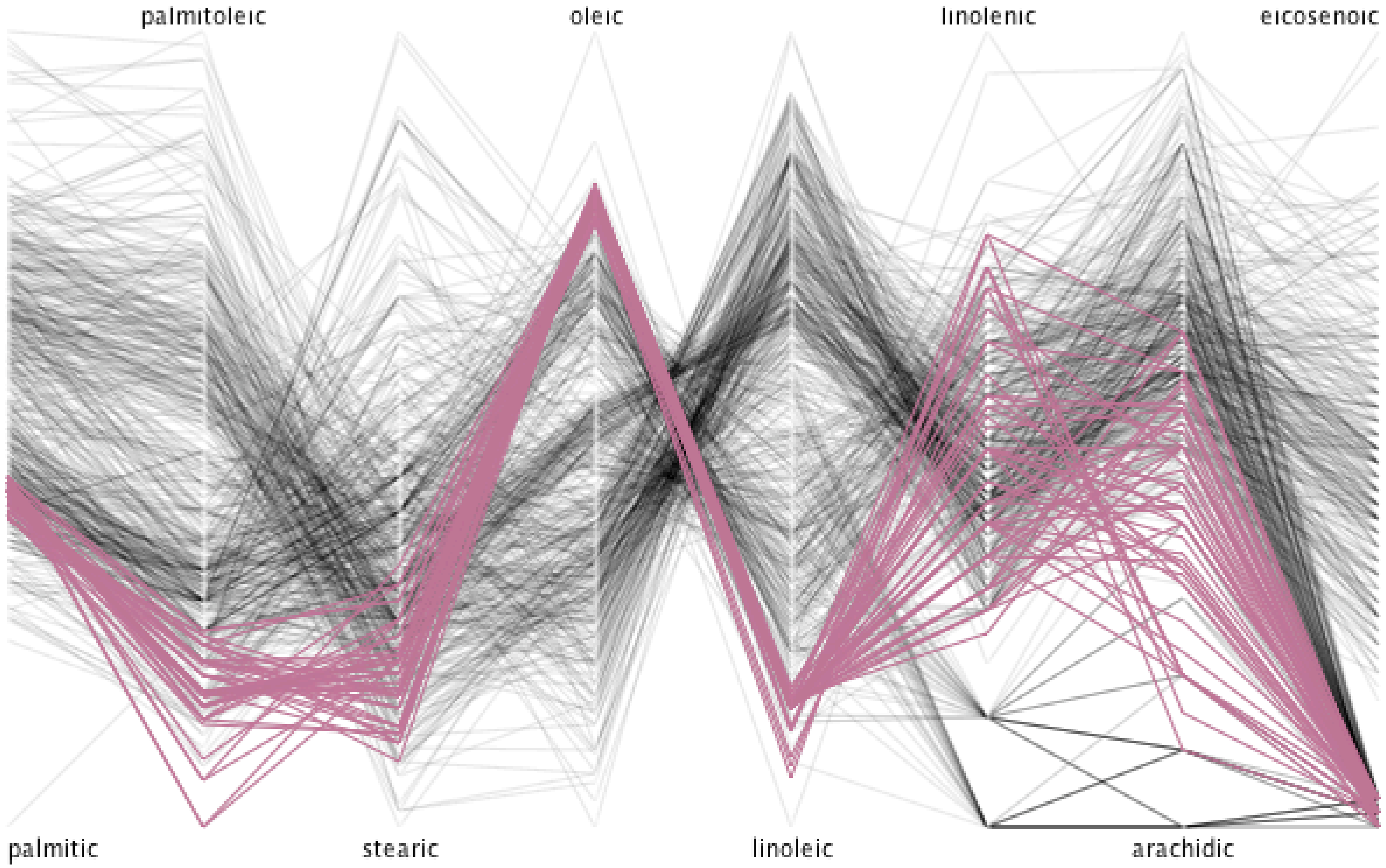
Spinograms: Properties

- Proportions can be compared directly
- x-scale is transformed → nonlinear, but continuous and isotone
- More visual weight on areas with high density
- There is a continuous counterpart using density estimators: CD-plot



Parallel Coordinates

- Overplotting is the most serious issue, but …
- … α-transparency is a good cure

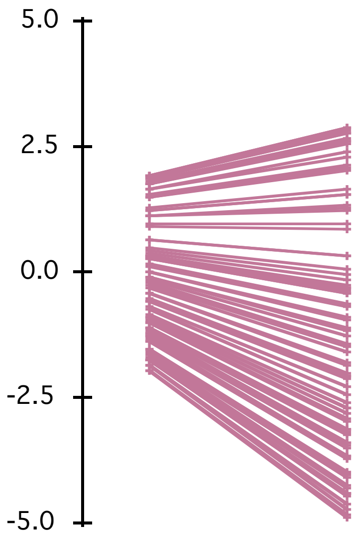


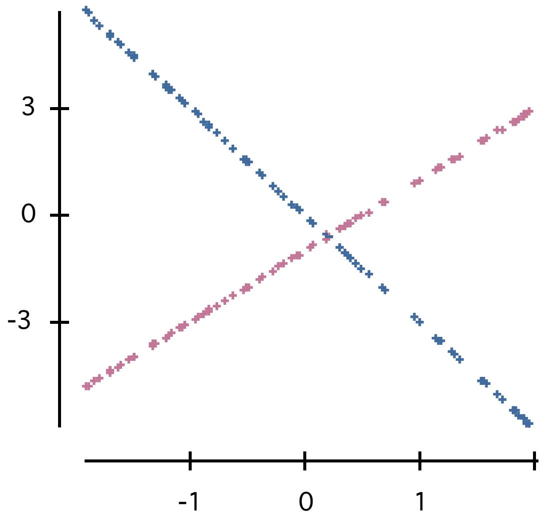
Parallel Coordinates: Ordering

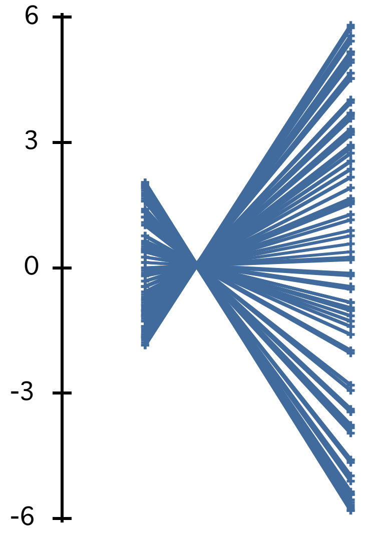
- Problem:
  We see interesting features best at adjacent axes

- For k variables we have $k!$ potential orderings.
  Each plot gives us $k-1$ adjacencies out of $\dfrac{k(k-1)}{2}$ potential adjacencies, leaving $\lfloor\dfrac{k+1}{2}\rfloor$ to look at.

- Sign of variable matters a lot



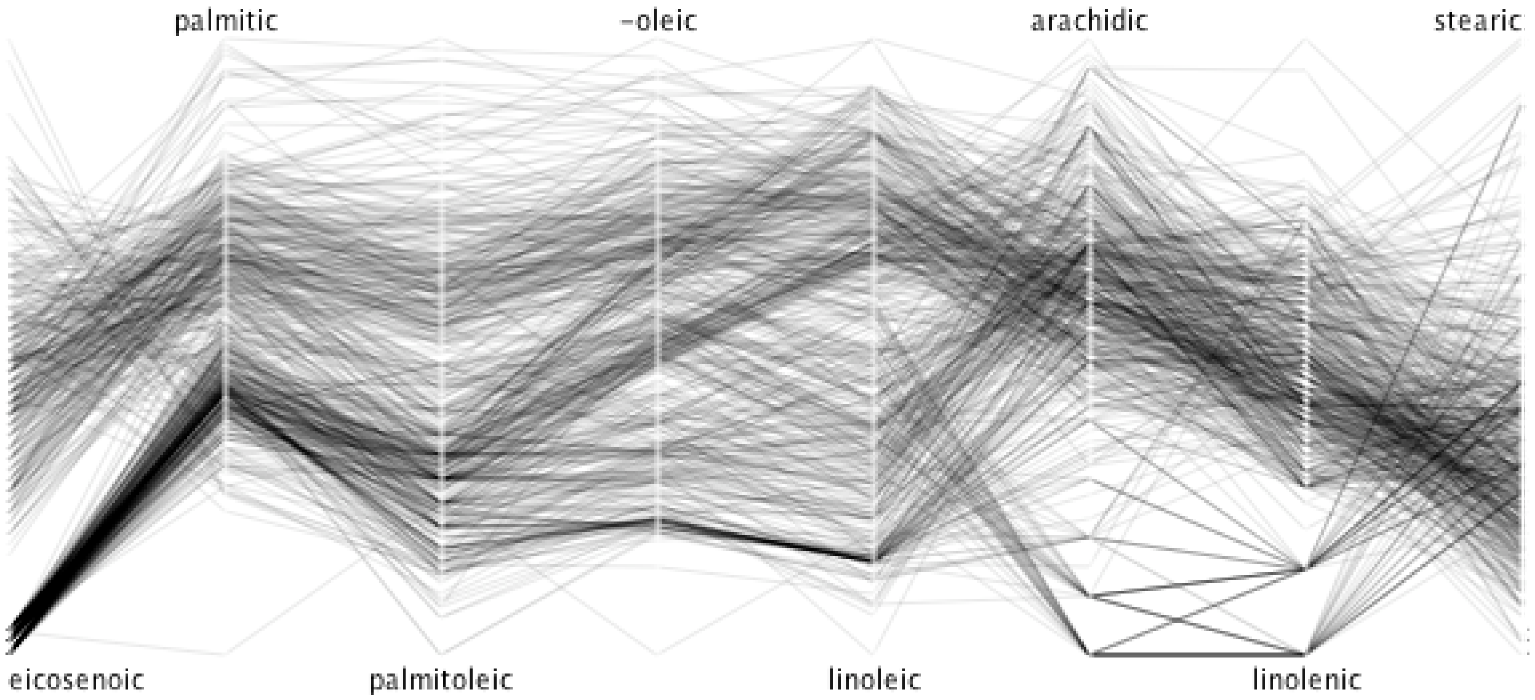




Parallel Coordinates: Correlations

- Correlations

```
             palmitic  palmitoleic     stearic      oleic    linoleic   linolenic   arachidic  eicosenoic
palmitic    1.0000000   0.83560497 -0.17039178 -0.8373354  0.46068446  0.31932669  0.22829912  0.50195179
palmitoleic 0.8356050   1.00000000 -0.22218545 -0.8524384  0.62162666  0.09311163  0.08548117  0.41635048
stearic    -0.1703918  -0.22218545  1.00000000  0.1135987 -0.19781693  0.01891719 -0.04097892  0.14037748
oleic      -0.8373354  -0.85243835  0.11359873  1.0000000 -0.85031837 -0.21817123 -0.31996234 -0.42414586
linoleic    0.4606845   0.62162666 -0.19781693 -0.8503184  1.00000000 -0.05743858  0.21097260  0.08904499
linolenic   0.3193267   0.09311163  0.01891719 -0.2181712 -0.05743858  1.00000000  0.62023577  0.57831851
arachidic   0.2282991   0.08548117 -0.04097892 -0.3199623  0.21097260  0.62023577  1.00000000  0.32866349
eicosenoic  0.5019518   0.41635048  0.14037748 -0.4241459  0.08904499  0.57831851  0.32866349  1.00000000
```

- 

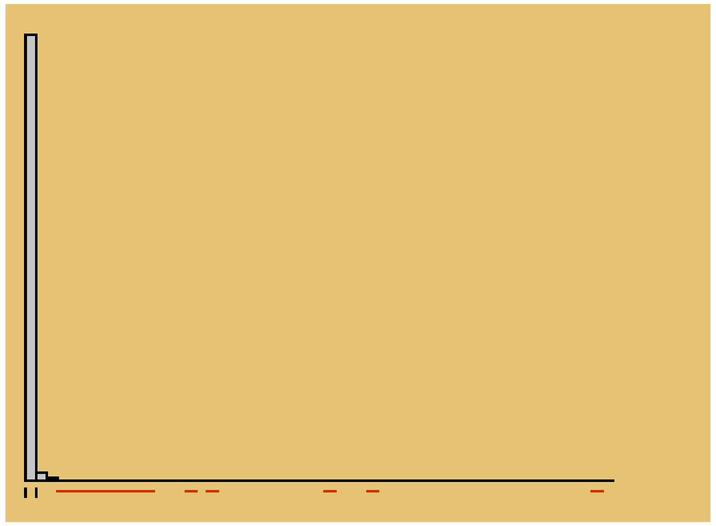
Parallel Coordinates: Summary

- Order matters
- Sign matters
- Simple orderings are:
  - min, max
  - range, var
  - mean, median
- Complex orderings à la Projection Pursuit Indices are more efficient
- … why not use projected data right away?

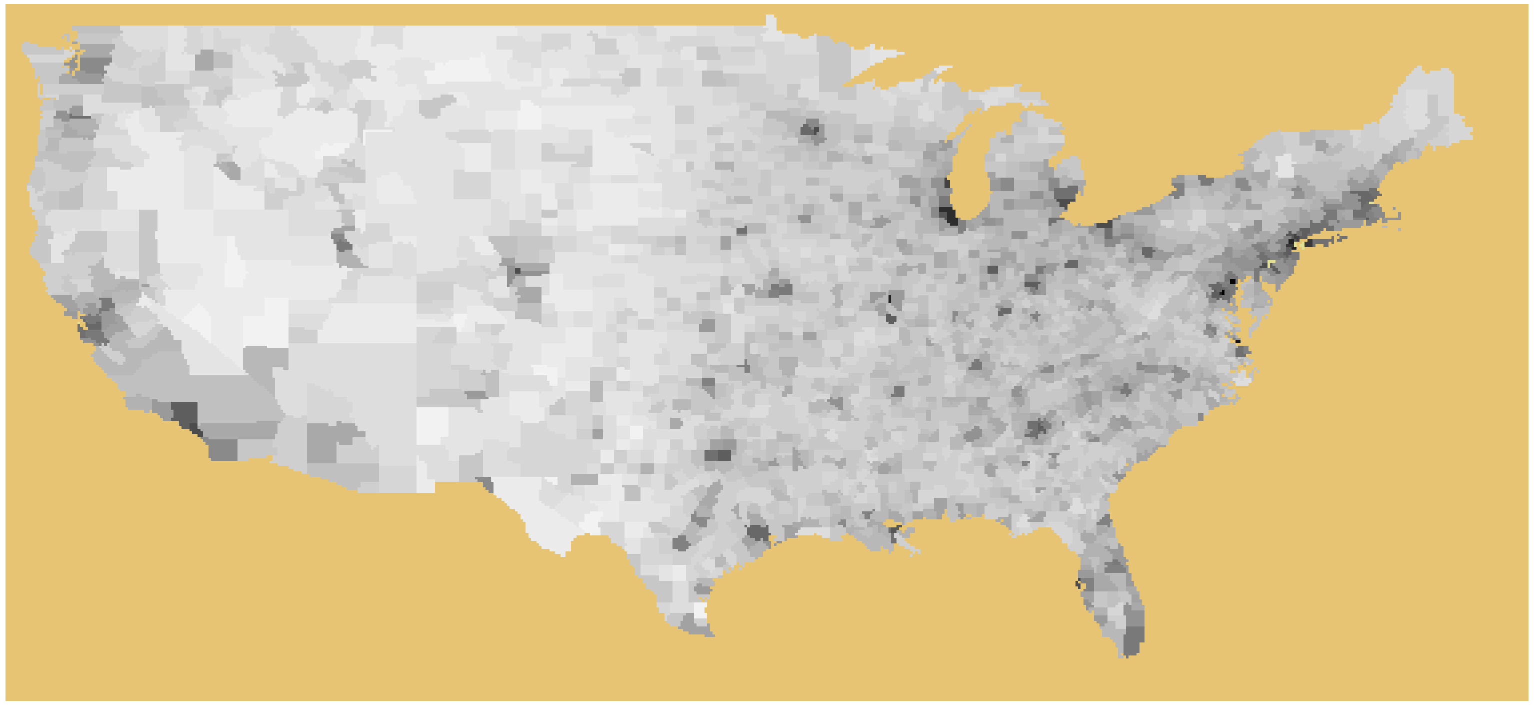
Choropleth Maps

- Static maps are usually very carefully optimized and fine tuned.
- Interactive environments need a fast and efficient implementation.
- Default: linear gray-scale is as simple as useless in most situations.
- Example: population density in 3072 US Counties.

Histogram



Map



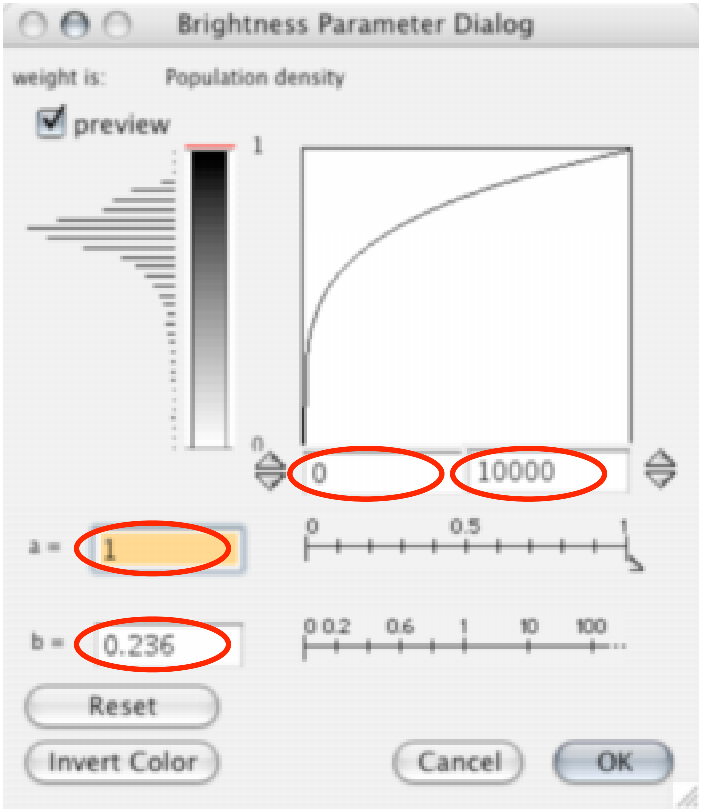
Choropleth Maps: Towards a better default

- Interactivity is very important, but still needs a good “seed”.
- MANET’s funktion to transform the grayscale:

$$f(x) := \begin{cases} a \cdot \left(\frac{x}{a}\right)^b & \text{for } x \leq a \\ 1-(1-a)\cdot\left(\frac{1-x}{1-a}\right)^b & \text{for } x \geq a \end{cases} \qquad \text{for } a \in [0,1] \text{ and } b$$

and dialog box:

- set a and b
- set max and min values of the mapping



Brightness Parameter Dialog

weight is: Population density

☑ preview

0 10000

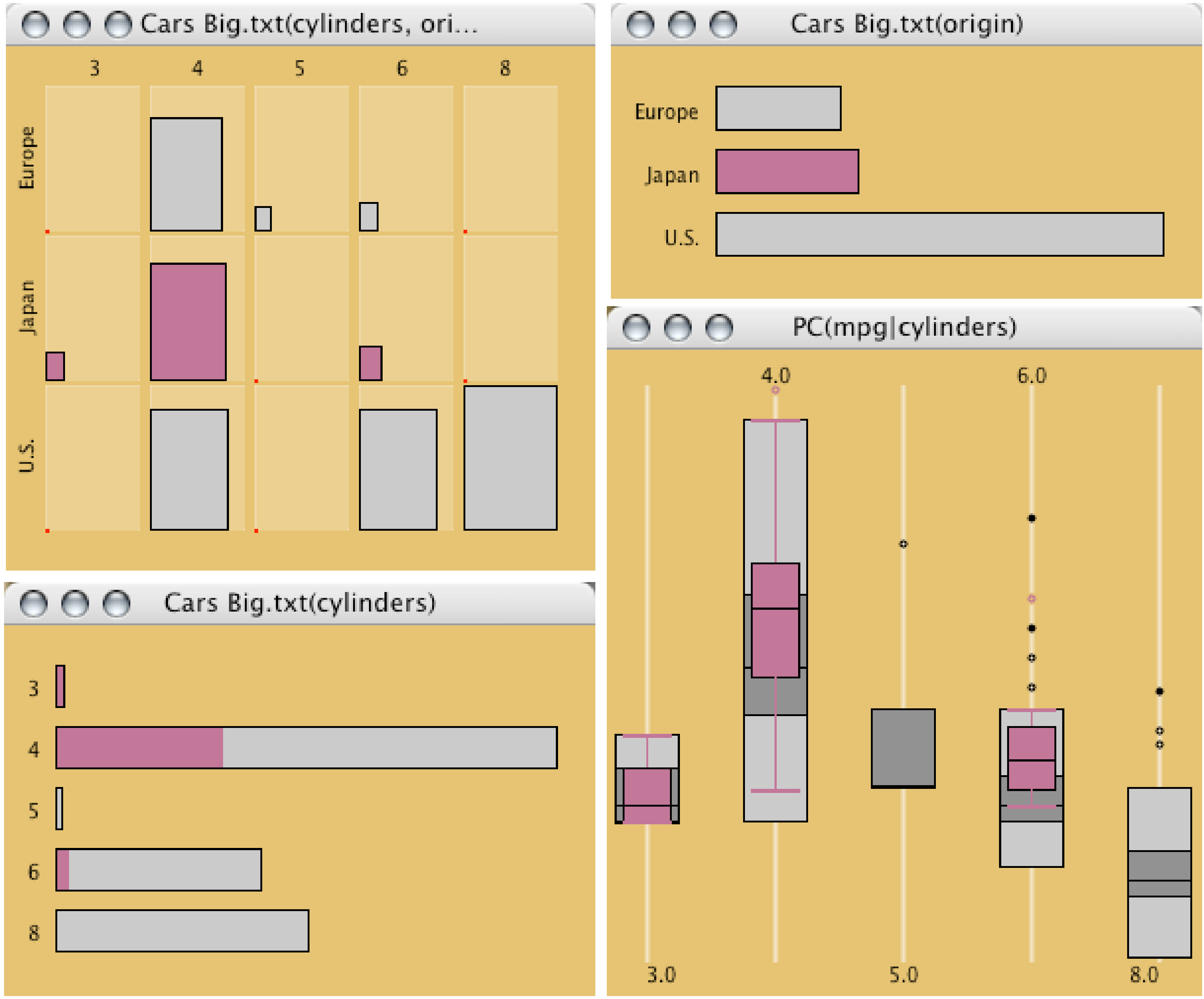
a = 1

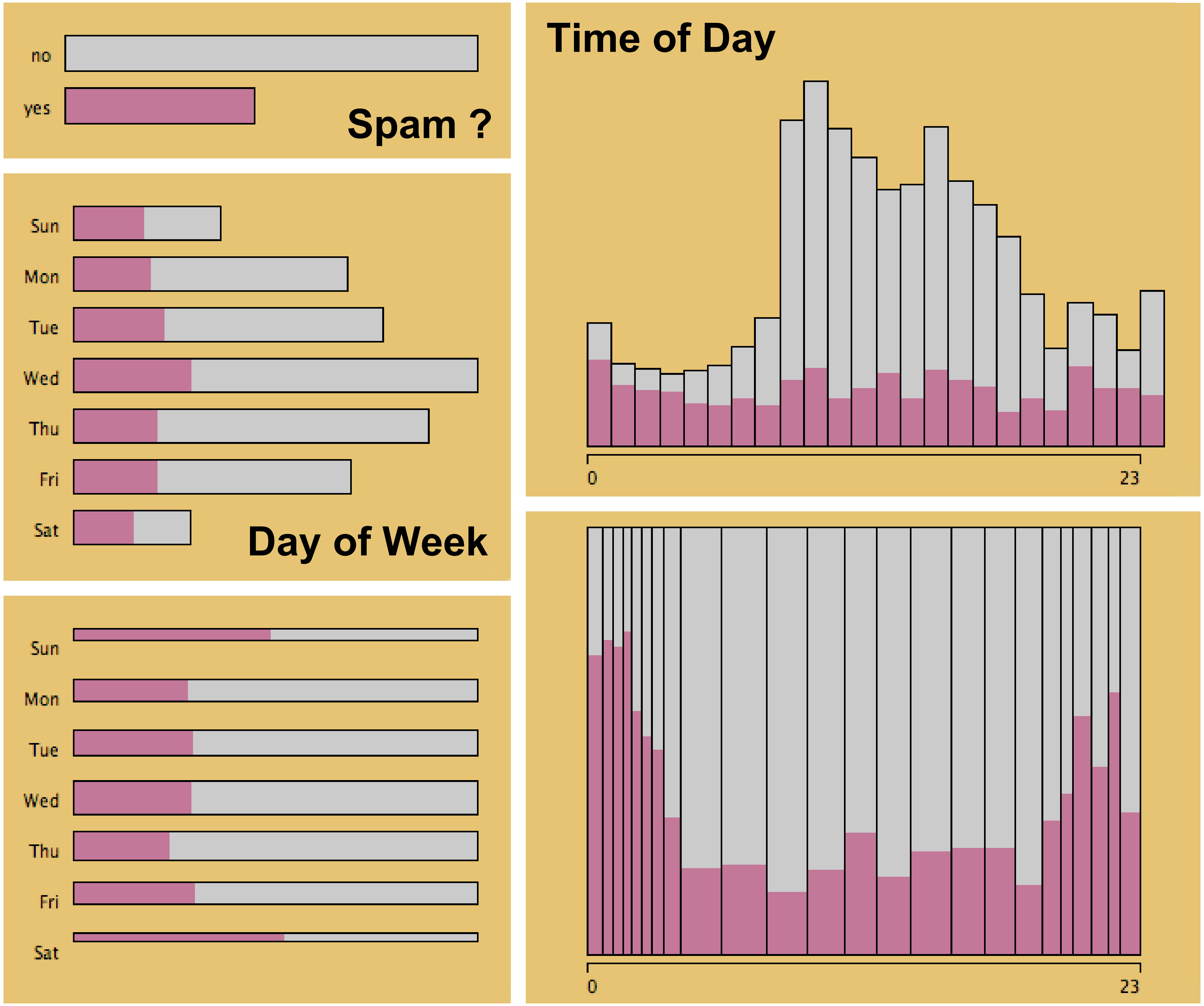
b = 0.236

- Easier:
  - Find suitable values for the four parameters to start off, i.e. which make the distribution of gray values “closer to normal”.

Plot Ensembles

- Idea:
  When looking at several variables simultaneously, certain combinations of plot types are especially useful.
- Example:
  Cars Data,
  "What is the influence of cylinder and origin on mpg?"





- “How are spam e-mails distributed over time?”

Conclusions

- Most defaults are not even sub-optimal
- Better rendering techniques are often helpful
- Rendering should adapt to the problem
- Intelligent orderings can be crucial
- Multivariate plot ensembles can guide an analysis

- Software should avoid plotting "chart junk"
- Defaults should adapt to what the user expects to get
  (I am not talking about MS's annoyances here …)

Thanks for your Attention