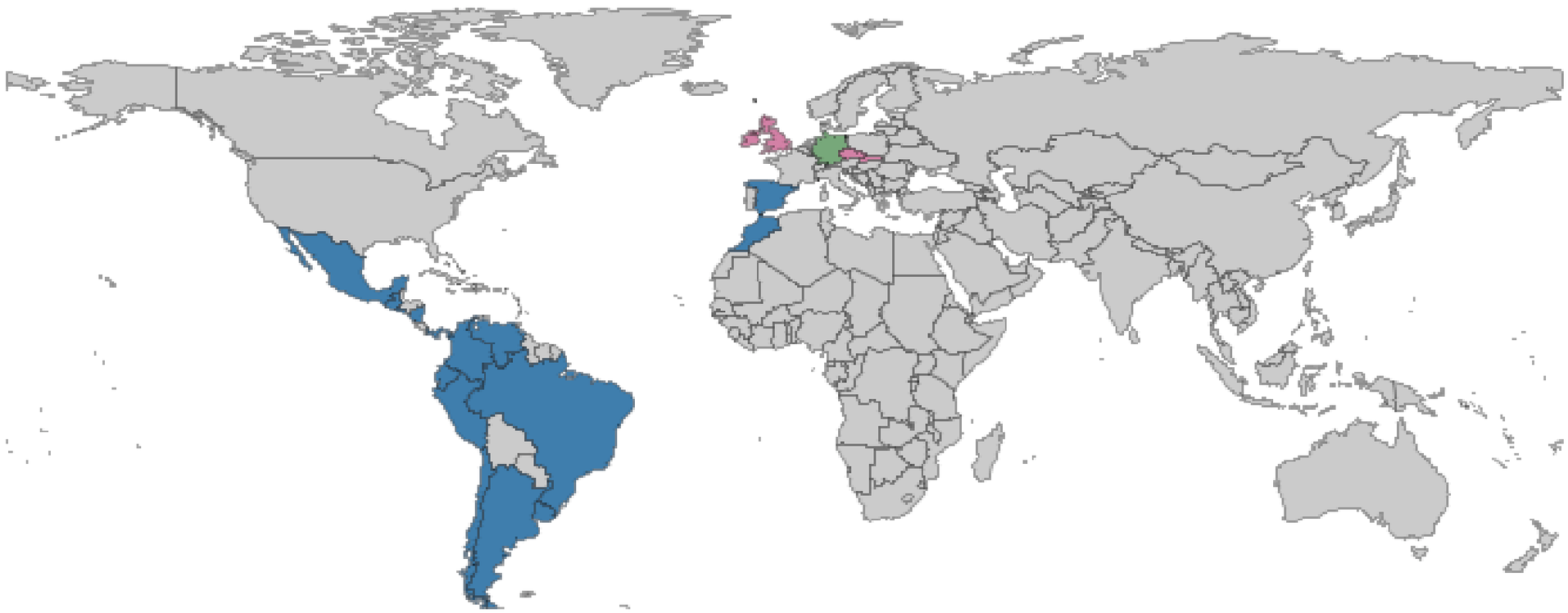
Statistical Graphics!
Who needs Visual Analytics?

martin@theusRus.de

Telefónica O2 Germany



Outline

- Data Visualization – who does not want to?
- Is statistical graphics more or less than InfoVis?
- From exploration to diagnostics and back?
- What have R Graphics and Susan Boyle in common?
- Where does R graphics head to?

Many Names – One Thing?

Information Graphics

Data Visualization

Info Vis

Visualization

Statistical Graphics

Visual Communication

Visual Data Mining

Visual Analytics

Many Names – Classification

Information

Data

Distributions

Information Graphics

Data Visualization

Info Vis

Statistical Graphics

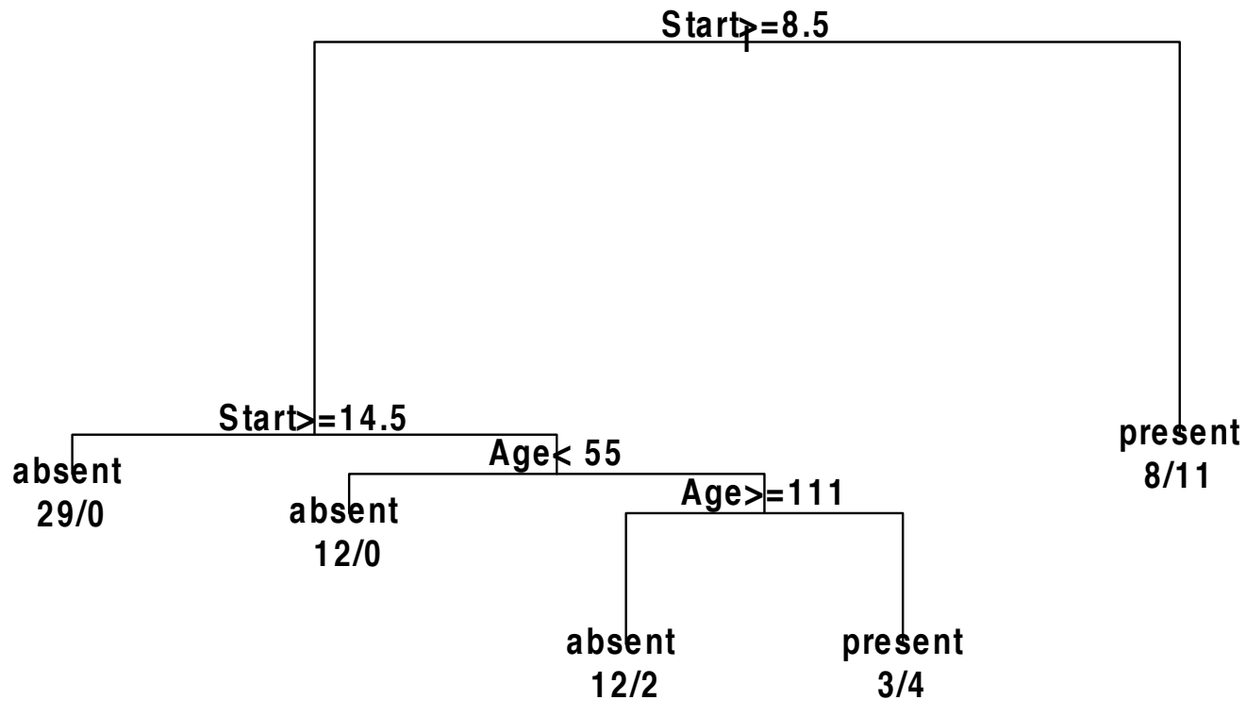
Visual Communication

Visual Data Mining

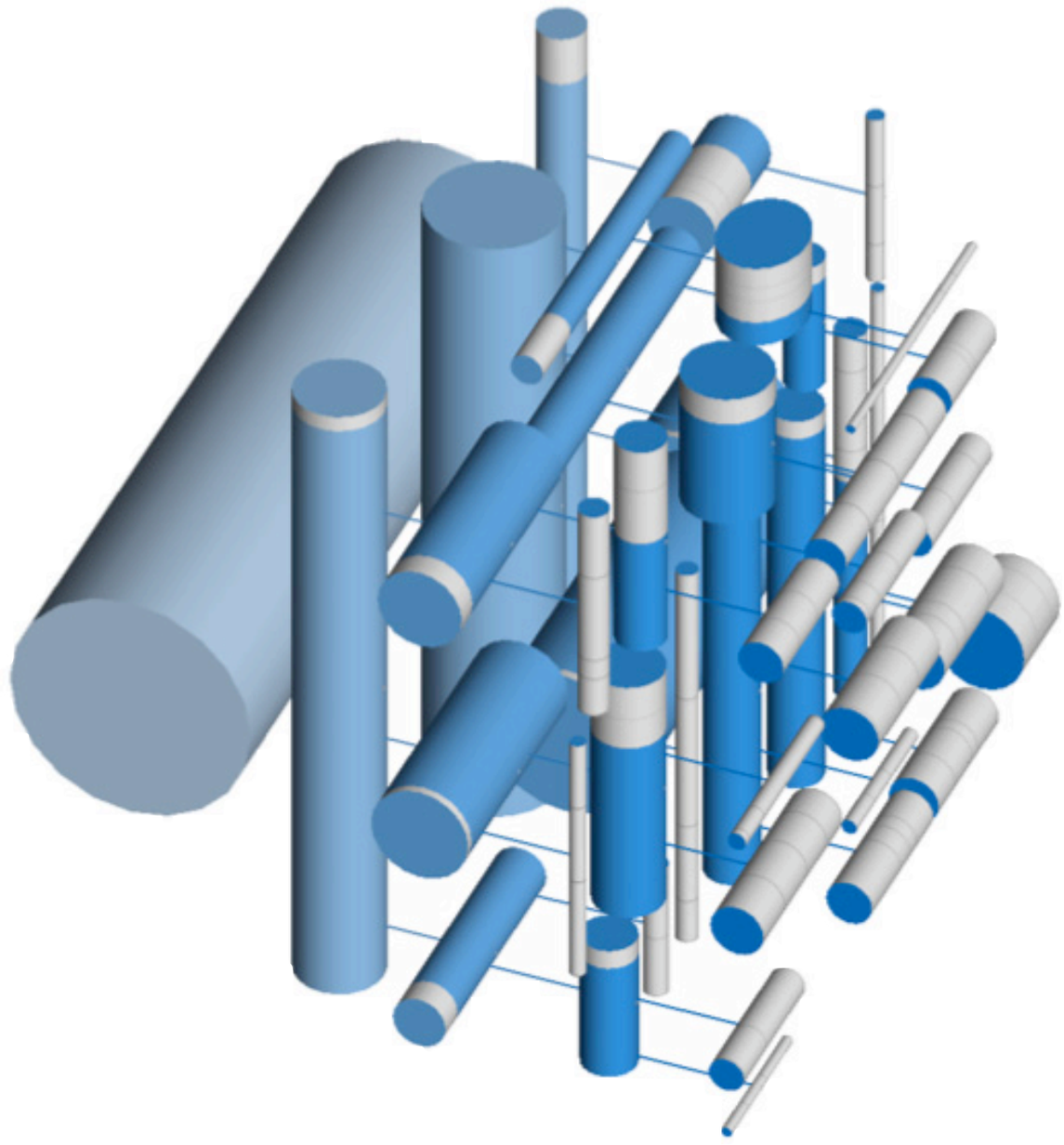
Visual Analytics

There are also Difference along other Dimensions ...

A Tree in R

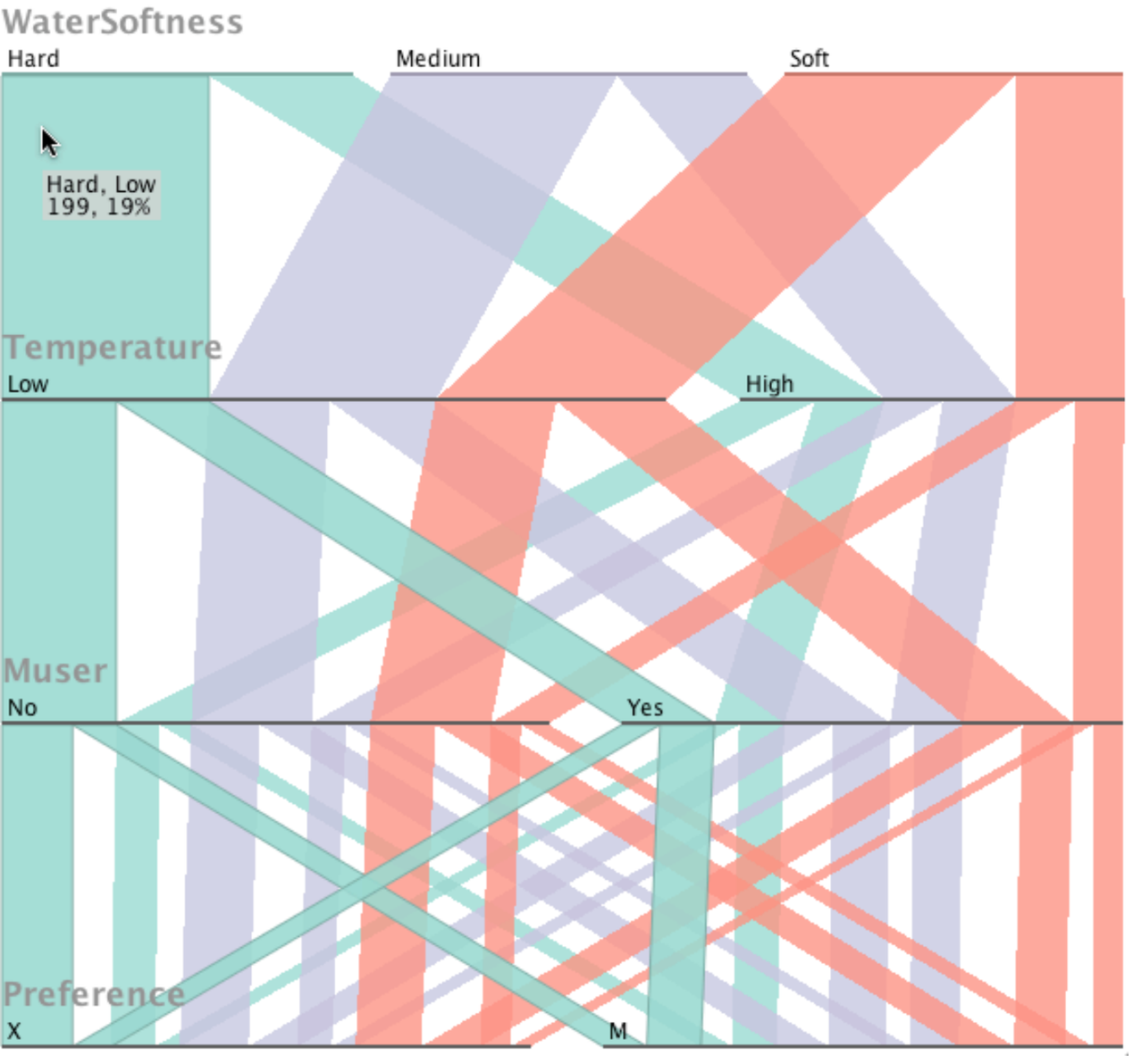


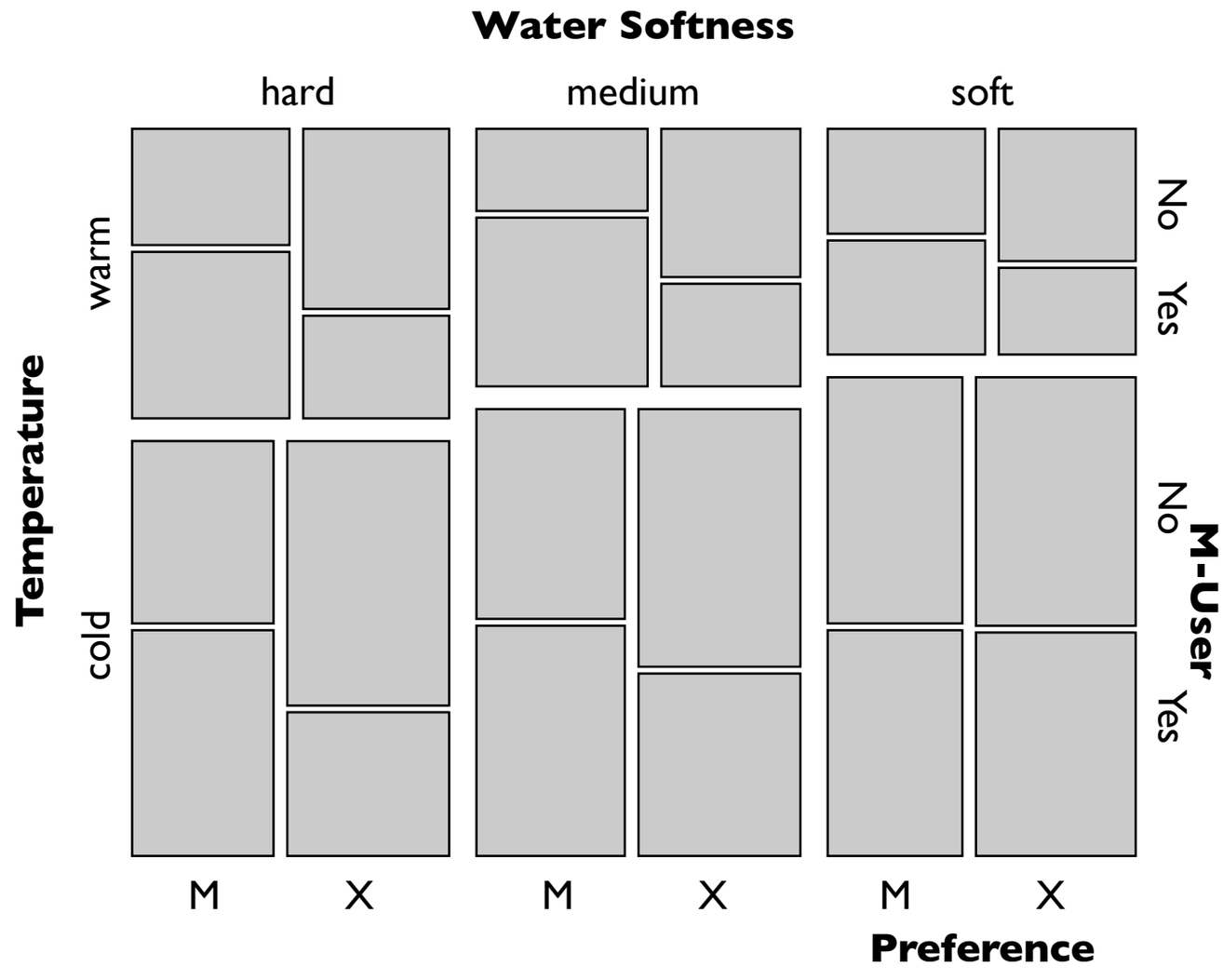
A Tree in Infovis



Where Statistical Graphics trials Visual Analytics I

- If a graphical display “only” shows the data it is much harder to go after certain properties we may expect to find in the data

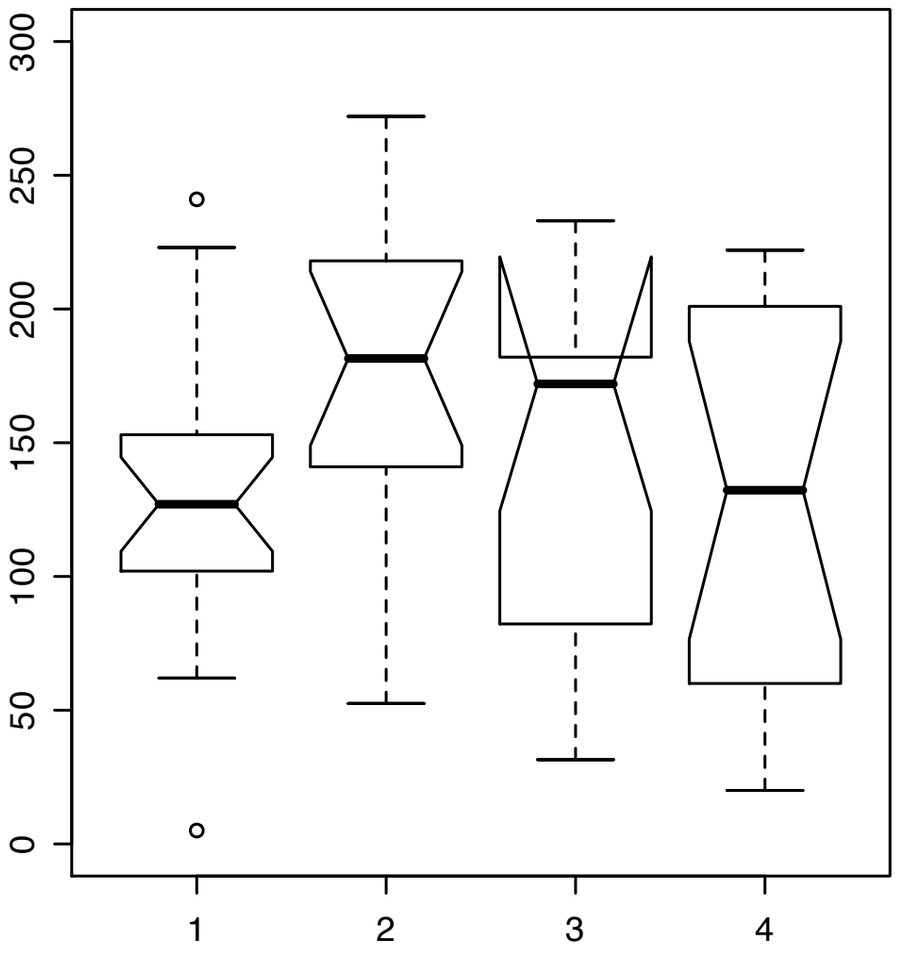




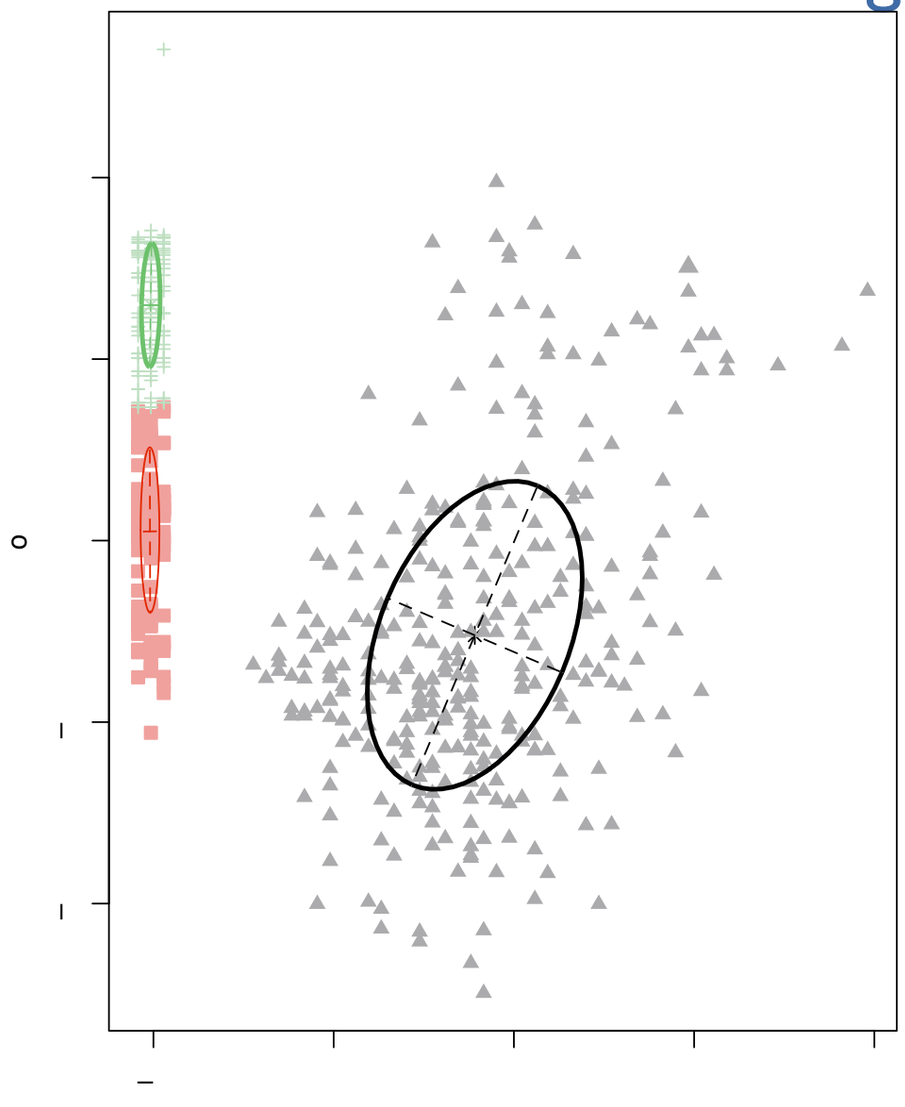
Where Statistical Graphics trials Visual Analytics II

- As a trained statistician we can look at graphics with distributions in mind ⇝ sometimes we add explizit decision support

Notched Boxplot

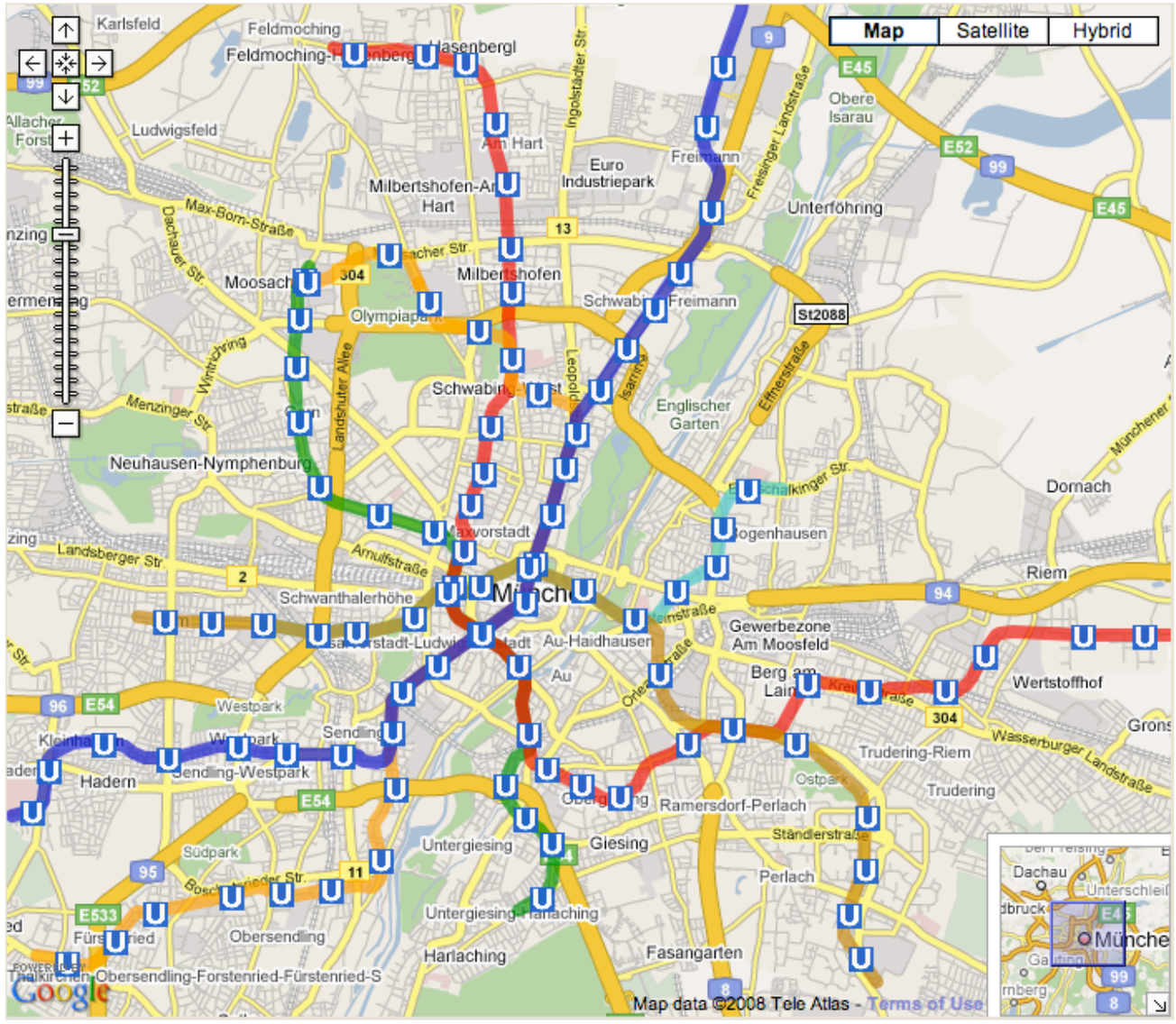


Model-based Clustering



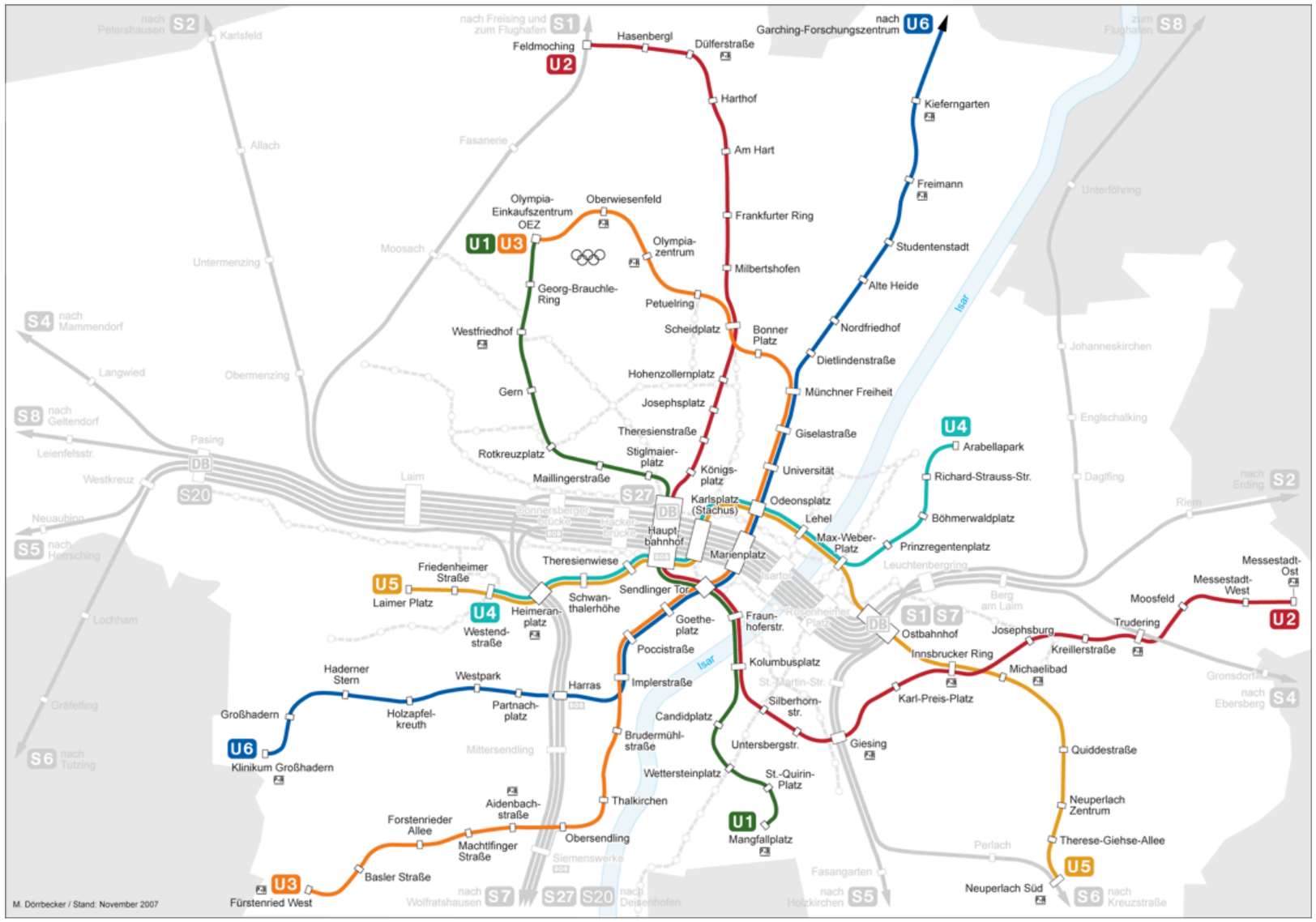
Visual Communication

- Common to all visualization efforts is to reduce the overall information to the relevant part that needs to be communicated



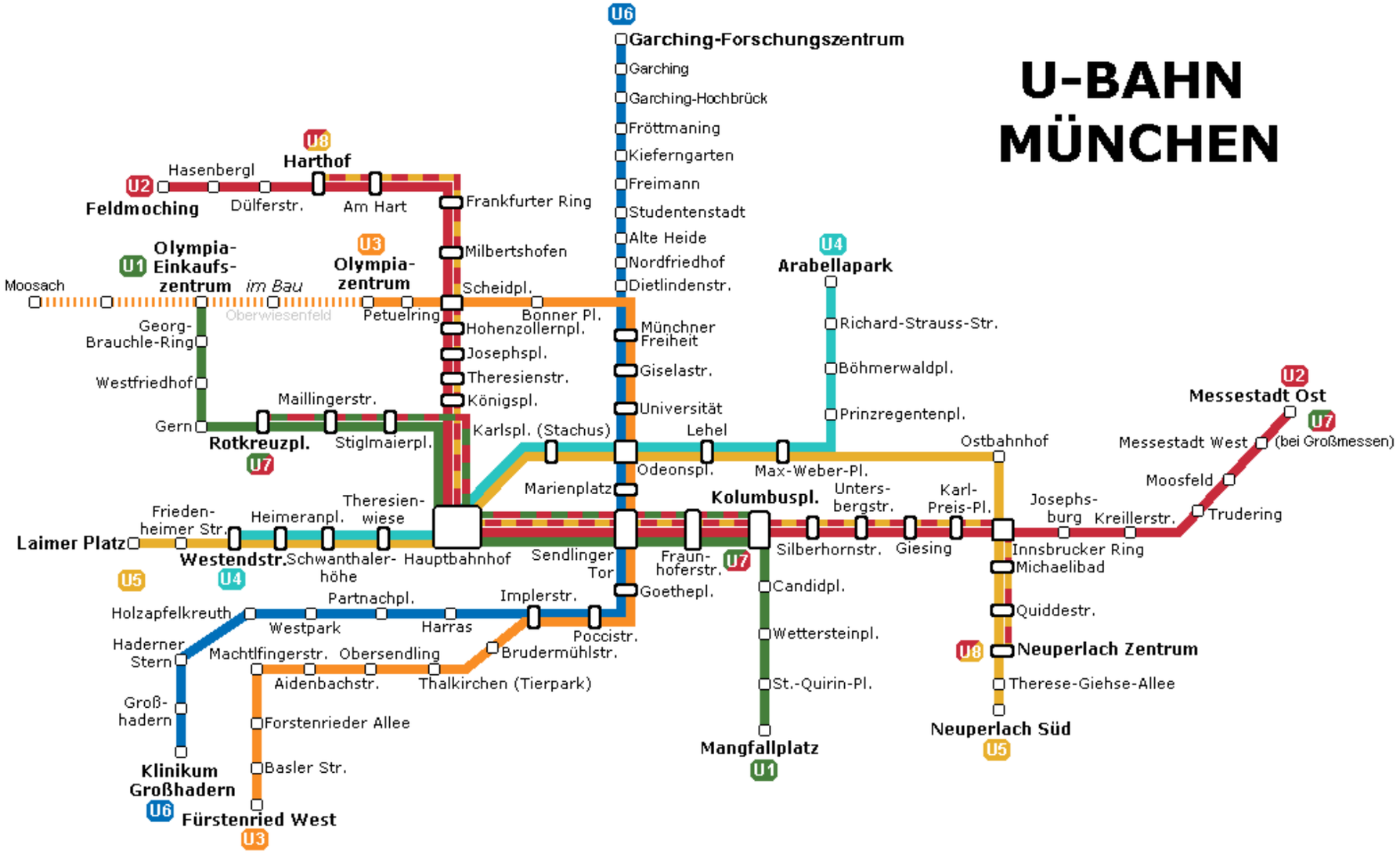
Visual Communication

- Common to all visualization efforts is to reduce the overall information to the relevant part that needs to be communicated



Visual Communication

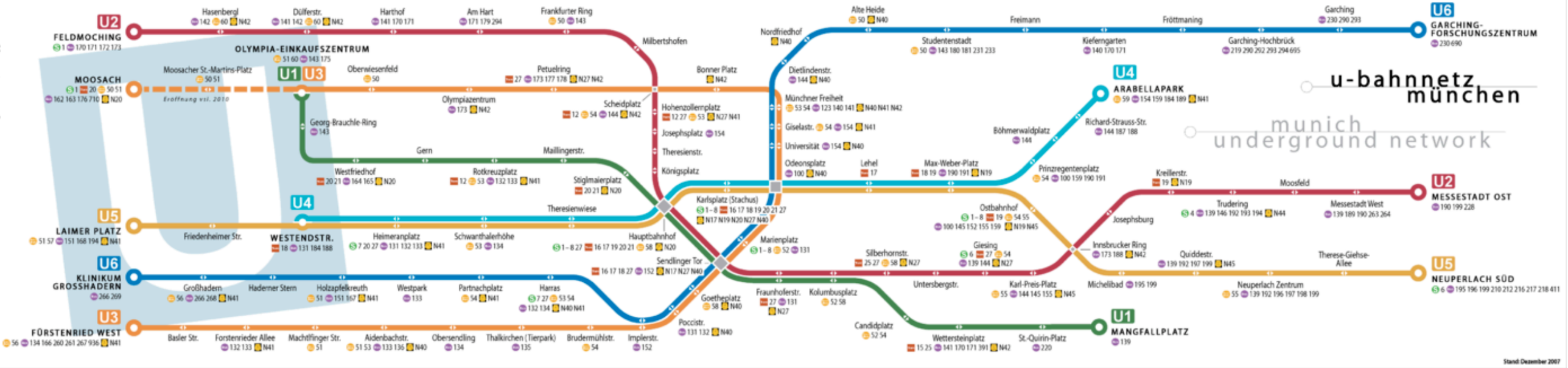
- Common to all visualization efforts is to reduce the overall information to the relevant part that needs to be communicated



U-BAHN MÜNCHEN

Visual Communication

- Common to all visualization efforts is to reduce the overall information to the relevant part that needs to be communicated



Constructing Information Visualization

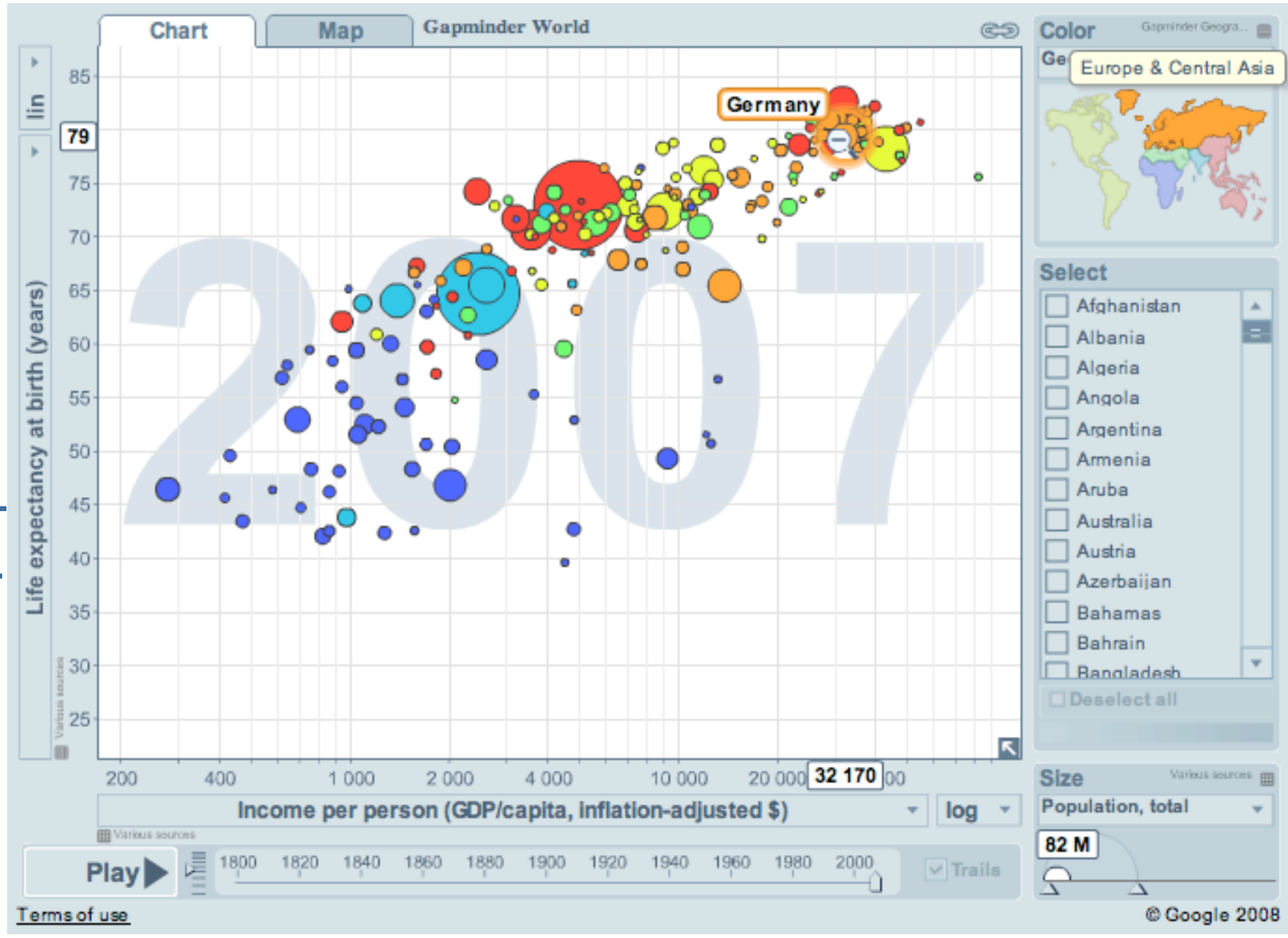
- Graphics design may be based on construction rules, design dogma or aesthetics, but all these points are neither necessary nor sufficient criteria for a successful design – but certainly a good point to start off.
- Milton Glaser puts it this way:

  "… *All design basically is a strange combination of the intelligence and the intuition, where the intelligence only takes you so far and than your intuition has to reconcile some of the logic in some peculiar way. …*"

- Is teaching good graph design then (almost) impossible?

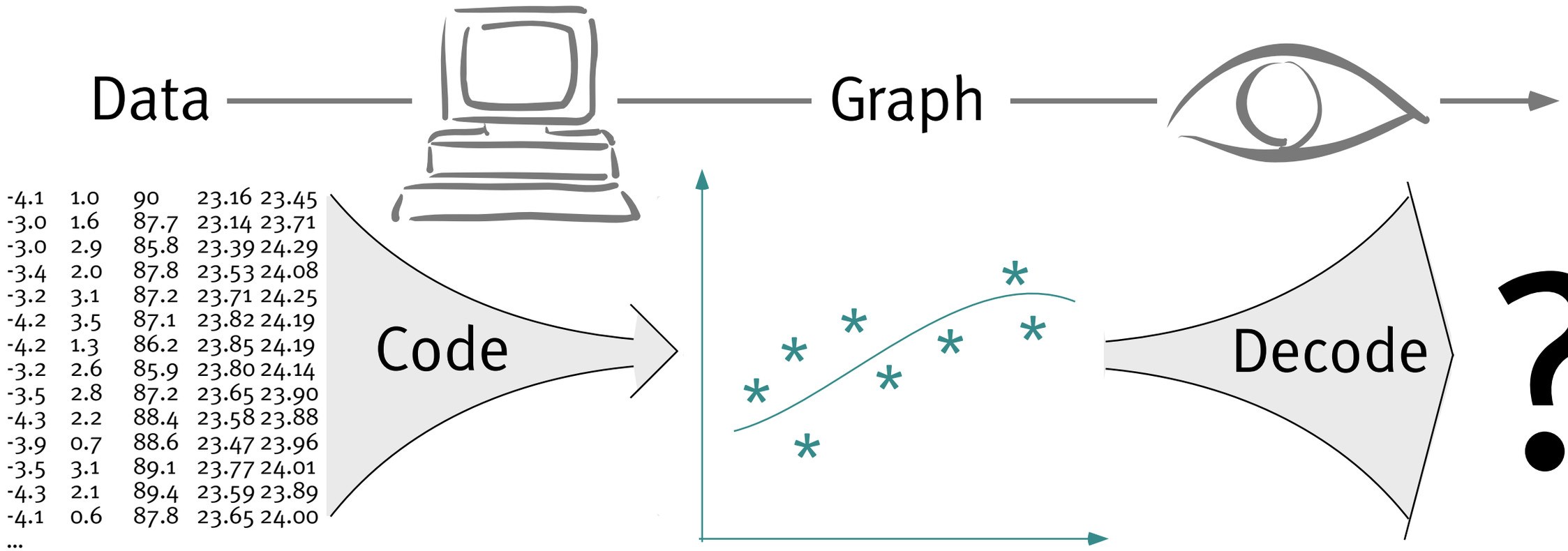
"Killer Applications"

- We ungrudgingly can confirm that we looked at multi[ple|variate] time series for quite some time, but the "narrative power" of the Gapminder animation is not met by any traditional display around
- Of course, the applications are still limited (three continuous measures for some dozons of categories) but in these cases they just work perfectly



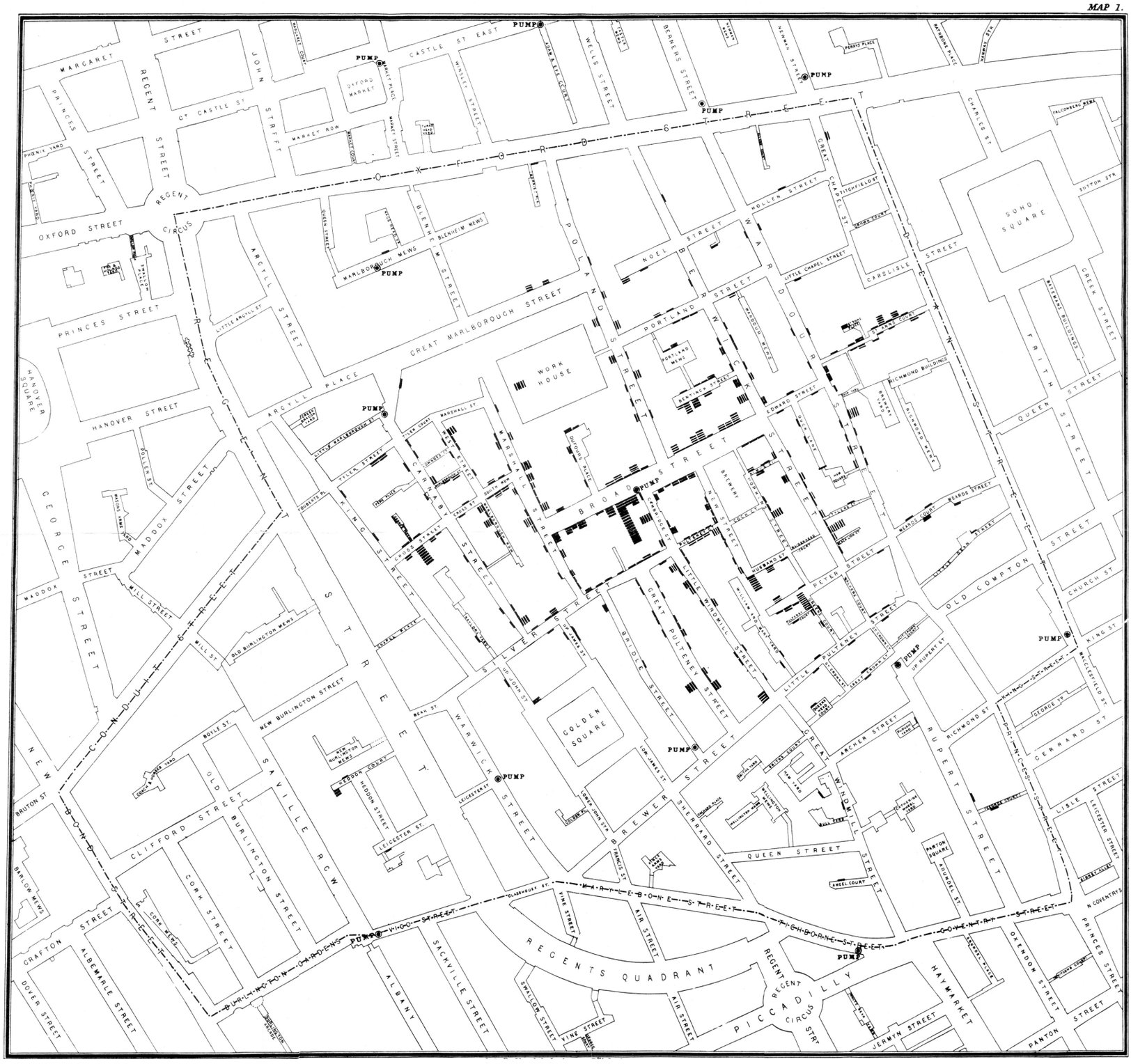
Processing Pipeline

- Perceptual aspects are central for the correct design and inter-pretation of graphics
- As perception is subjective, we need to get it as unambiguous as possible
- Reusing common building blocks eases the decoding of graphics



Drawing immediate conclusions from graphs

- Usually graphs are far weaker in communicating precise information than tables, such that the surplus of graphics must be the qualitative take home

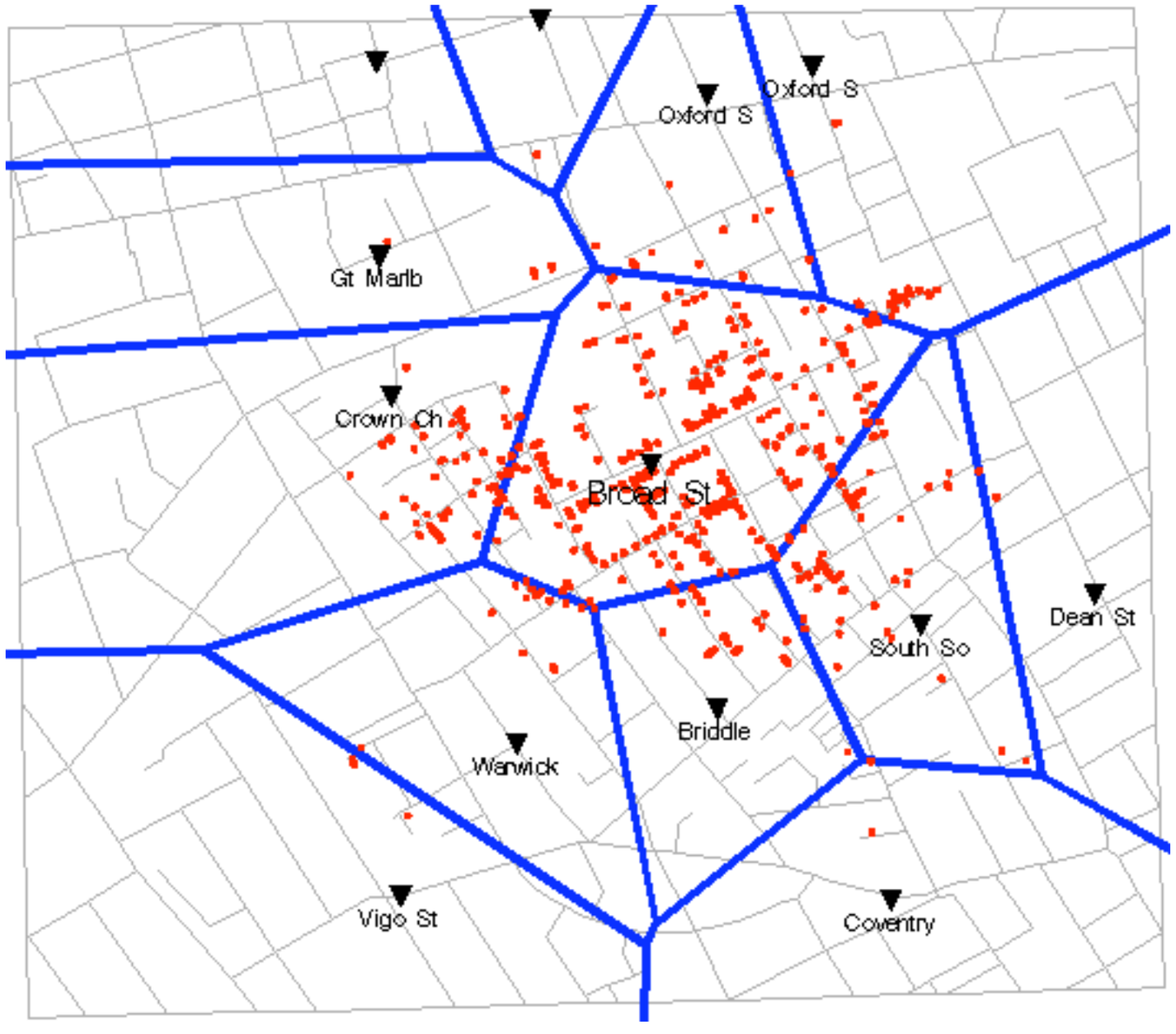


Dr. Snow's cholera desease map from 1855

Snow's Map: Visual Processing Pipeline

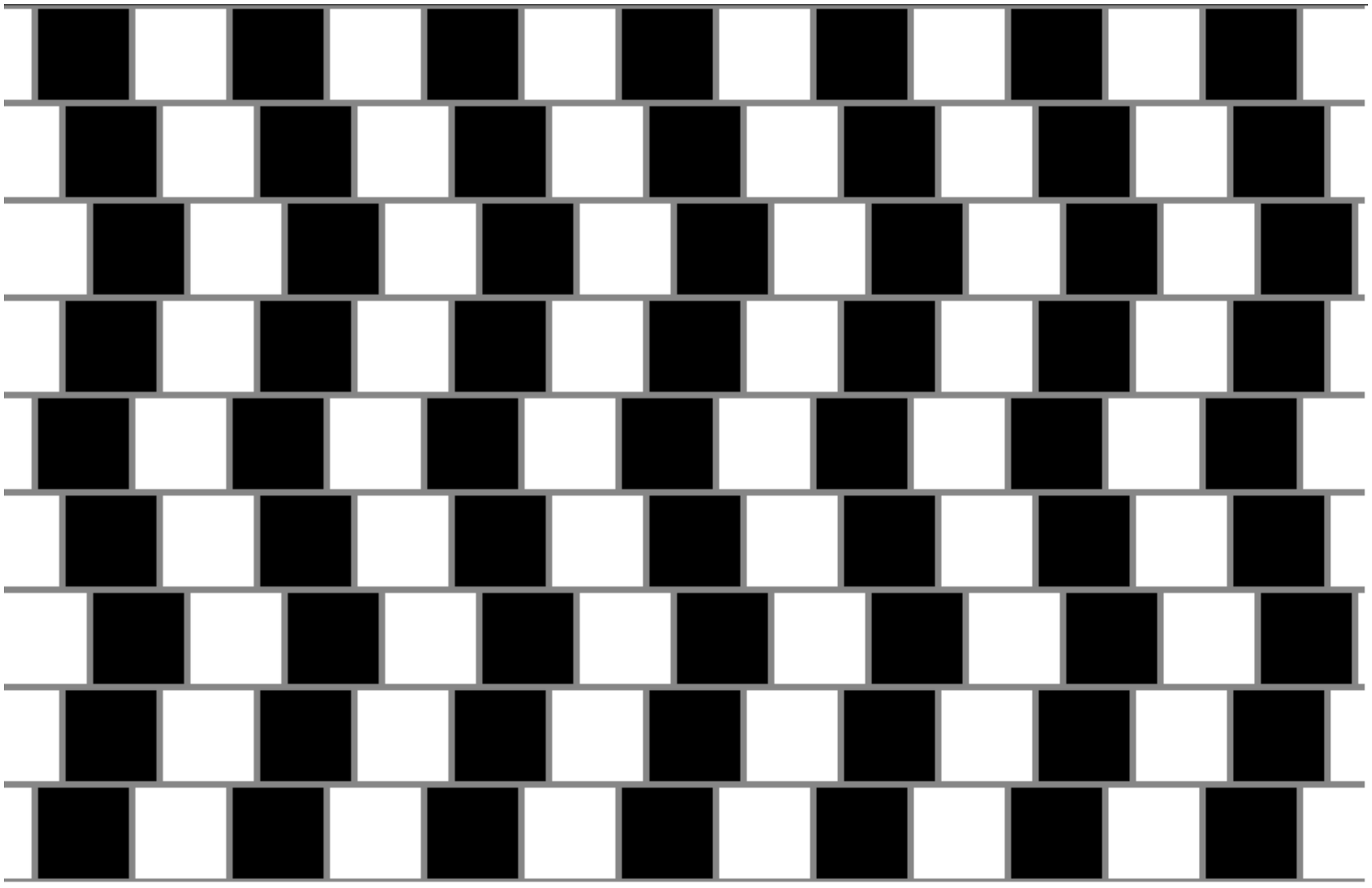
How do we learn from the map?

Three Steps
1. Mapping Cases
2. Mapping Pumps
3. Judging Distances



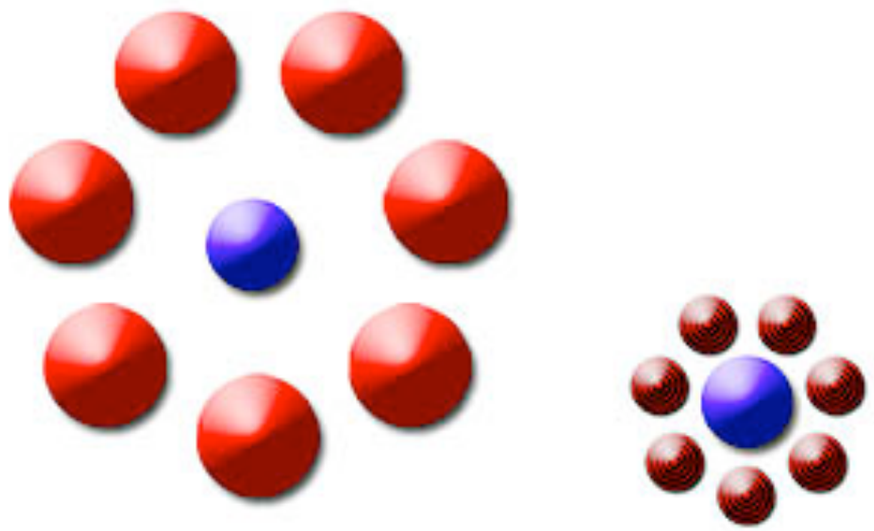
Some Eye Candy ...

parallel lines

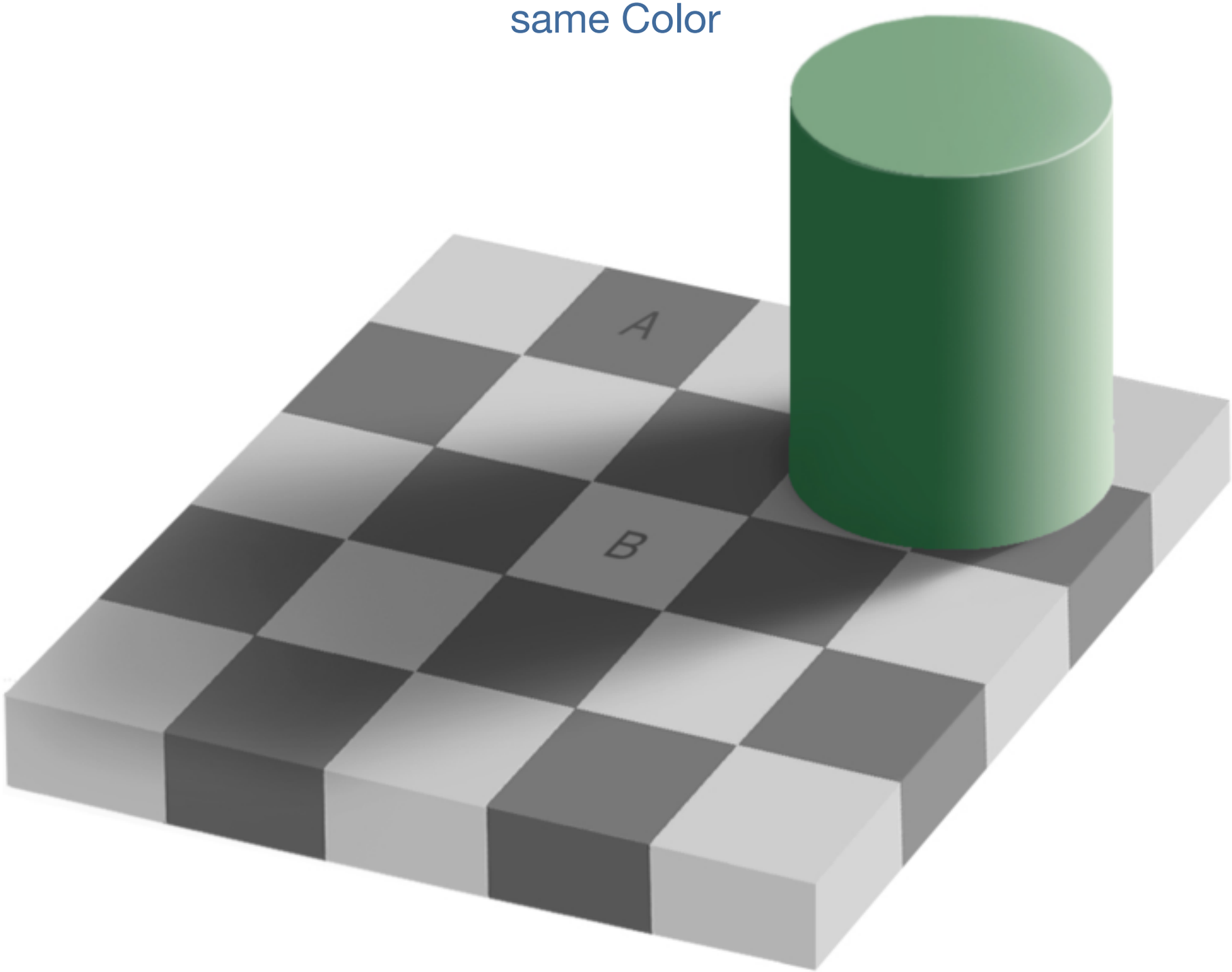


Some Eye Candy ...

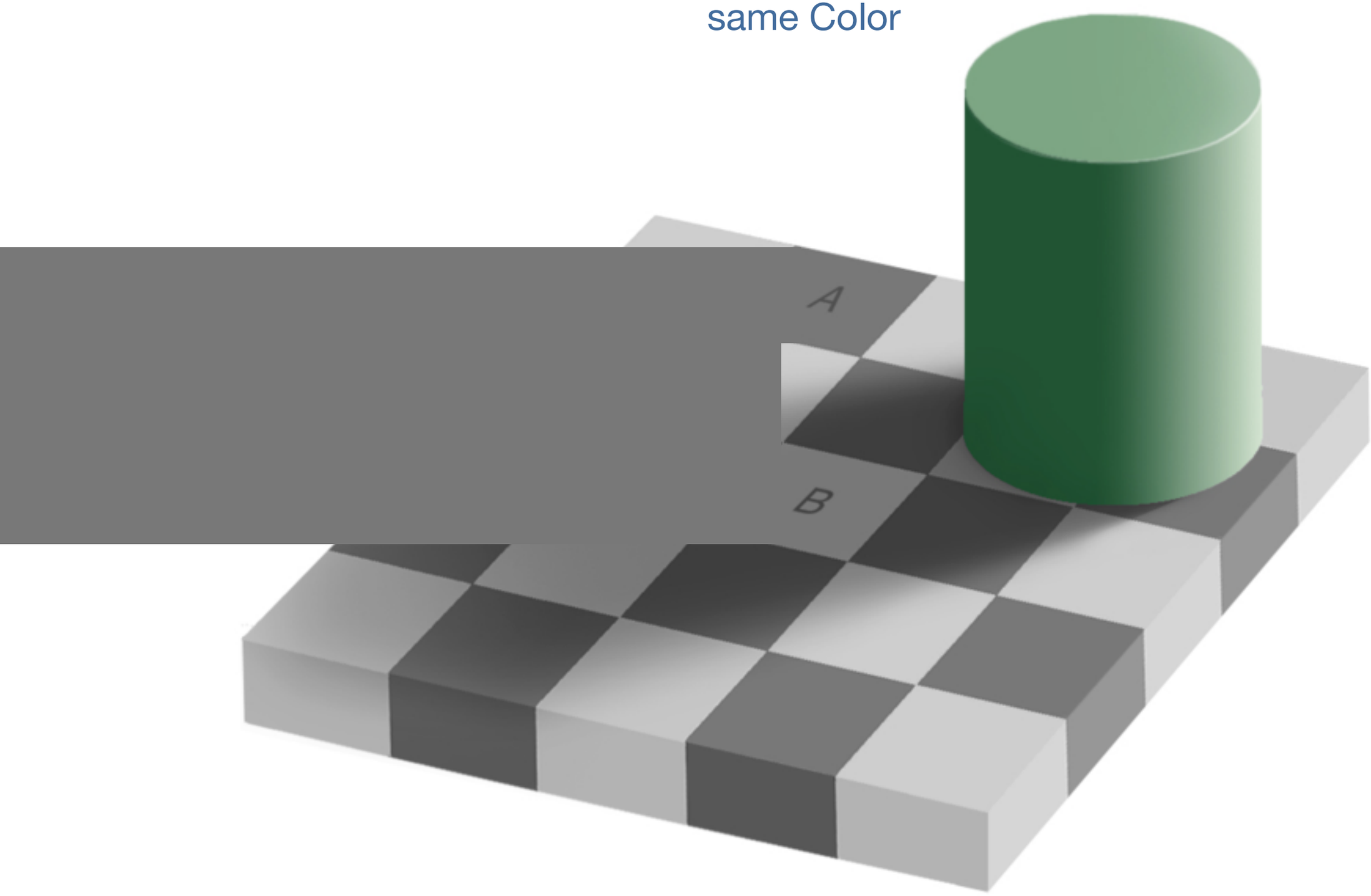
same Size



Some Eye Candy ...

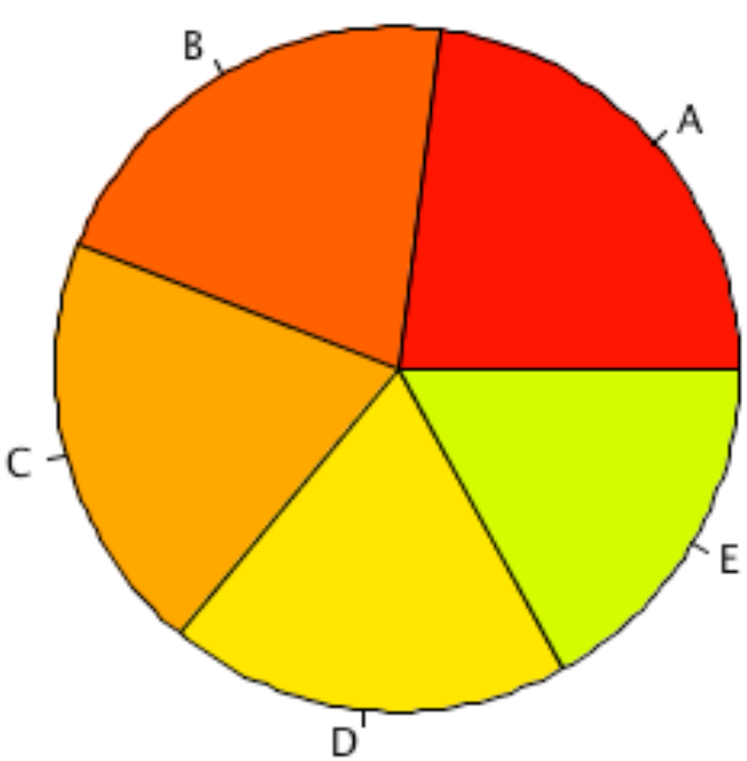


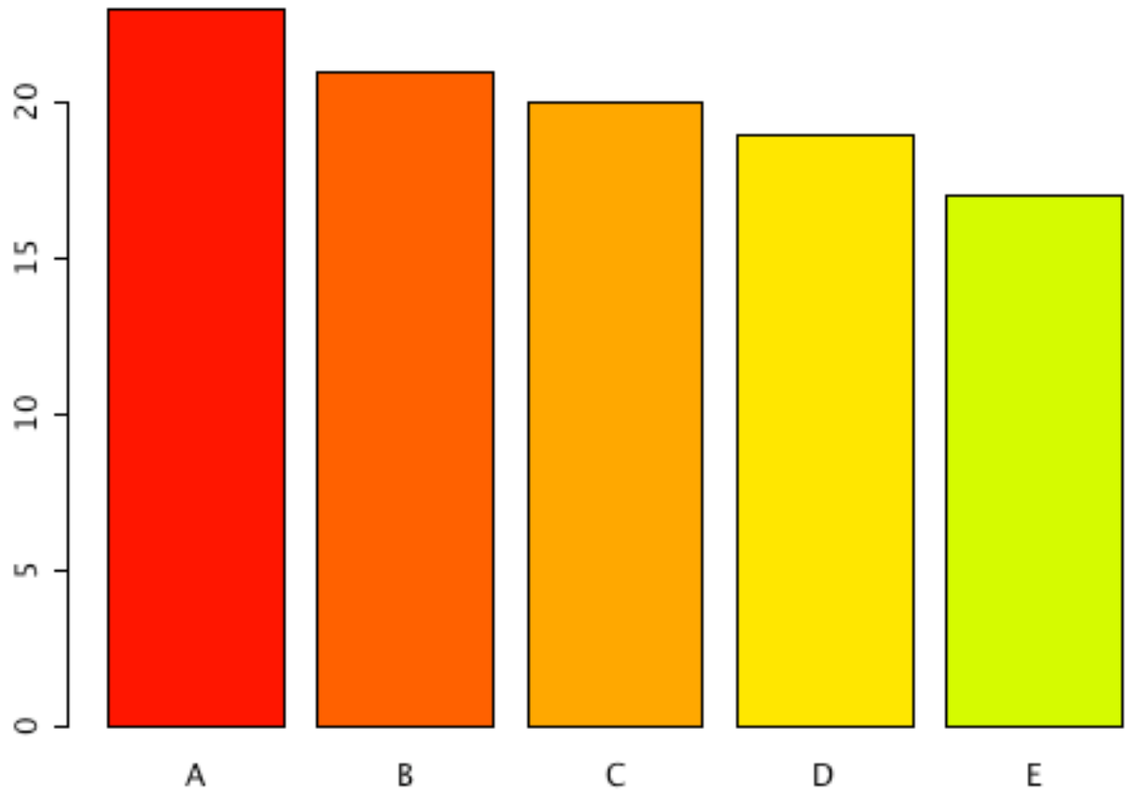
Some Eye Candy ...



The eye isn't equally good at judging different shapes

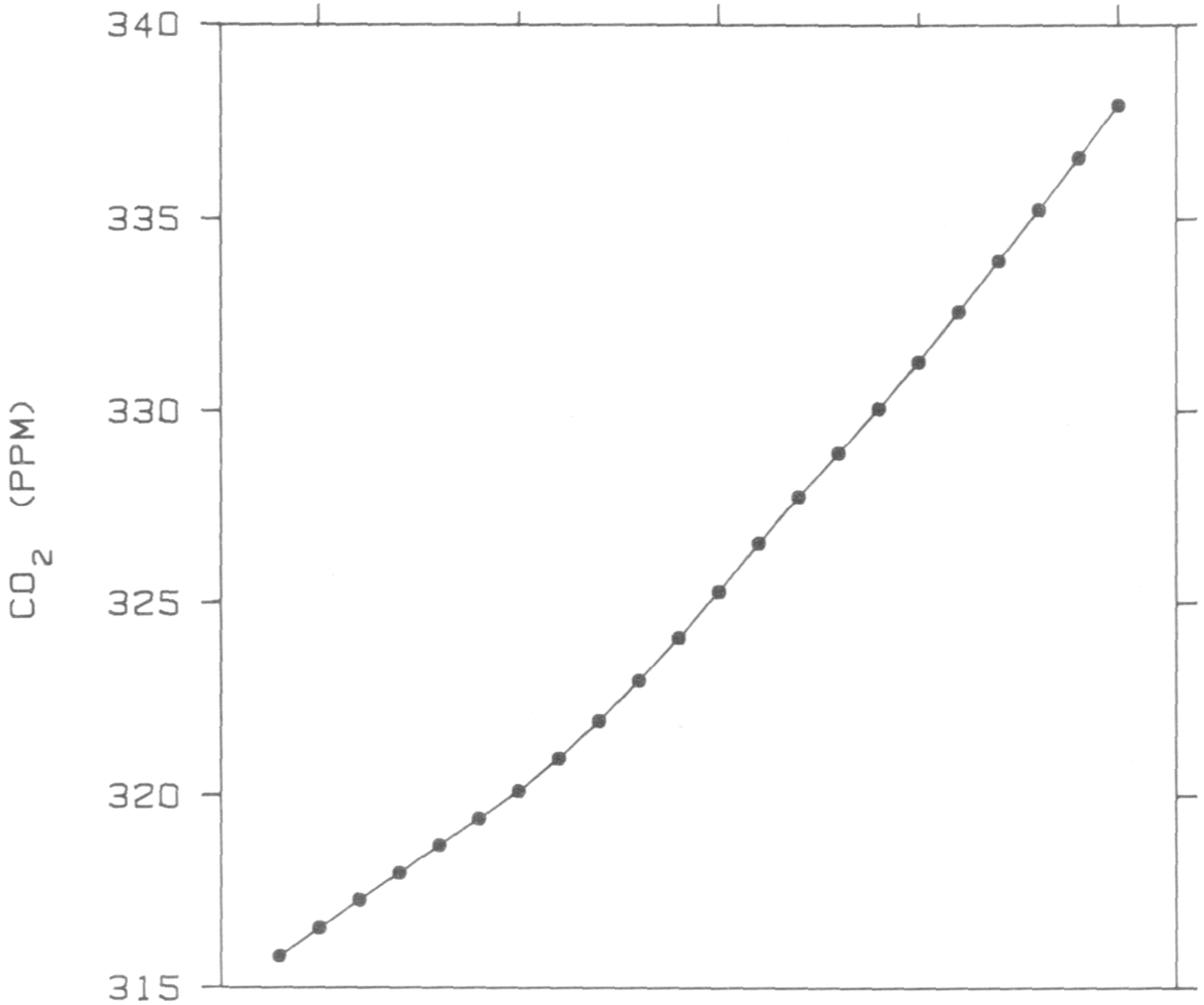
- Angles are much harder than distances …



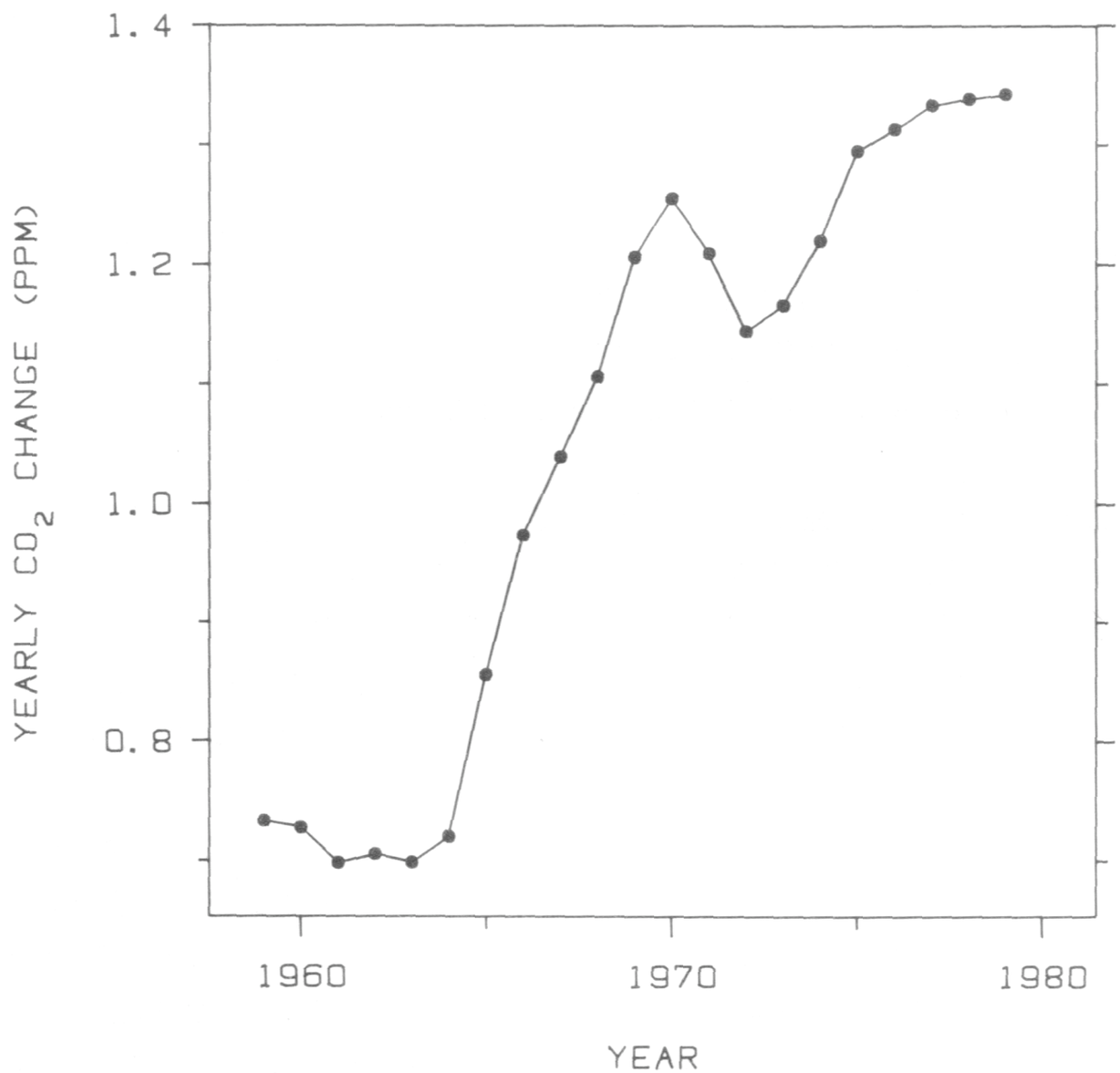


The wrong plot might obscure the message

The graph shows yearly CO_2 concentrations.
What can you tell about the slope?



The first differences (year to year change) shows surprising but not quite reasonable results.



Building Blocks for Statistical Graphics

- **Points**
  Points usually represent single observations, which are put along scales into a coordinate system.

- **Rectangles**
  Rectangles usually represent a group of observations, and the size of the rectangle should be proportional to the number of observations in this group.

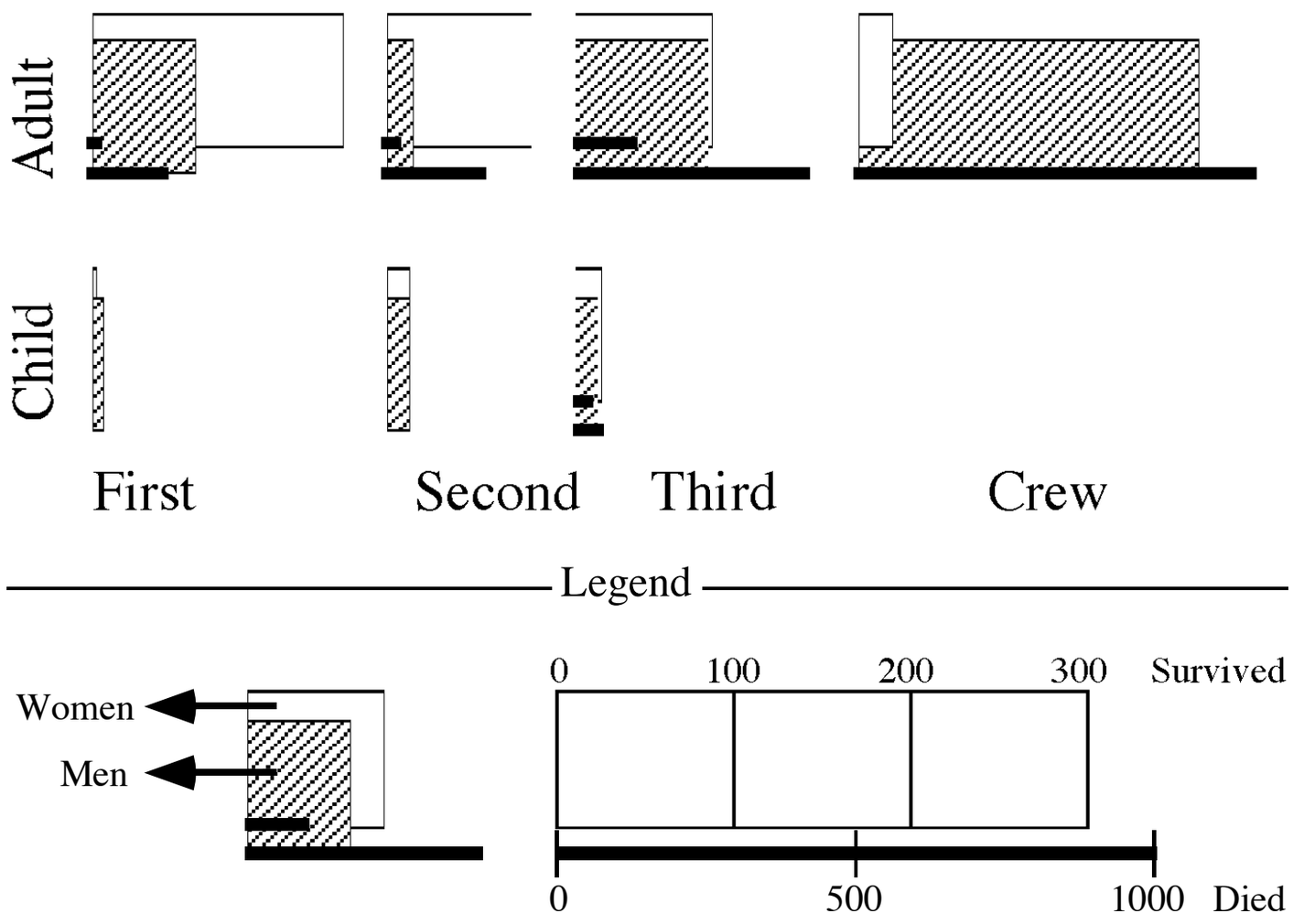
- **Lines and Polygons**
  Lines are usually used to join depend observations of the same entity, i.e. one polyline represents one entity.
  Polygons (like in maps) are usually used as a generalization of rectangles and used alike.

Consistent use of Graphical Primitives is a Must

- Bertin's (1967) proposal of displaying multidimensional discrete data.

- The data here:
  Titanic Passengers
  - Class
  - Age
  - Gender
  - Survived

- Can you tell "the Story"?
  (from the graph,
  NOT from the movie)

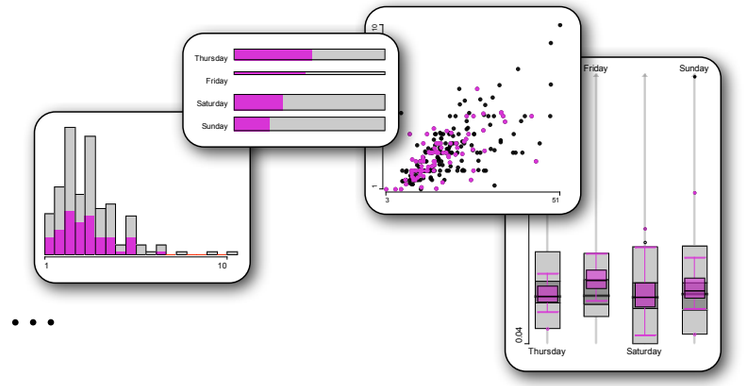


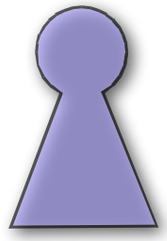
I any case we should get the fundament(al)s right!

- Use “the right” building blocks
- Make sure to use scales appropriately
- Adhere to standards when possible, be creative when necessary
- Seek generality as much as possible

The Use of Graphics: Part I

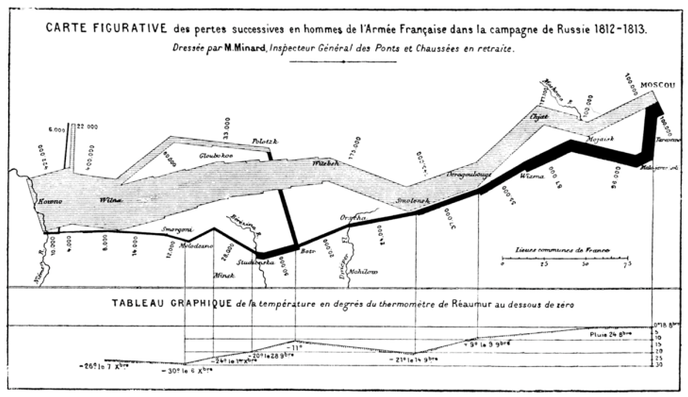
- Exploration
  - Aims at gaining insights
  - Mainly personal use
  - Few scales and legends
  - Highly interactive and little persistent

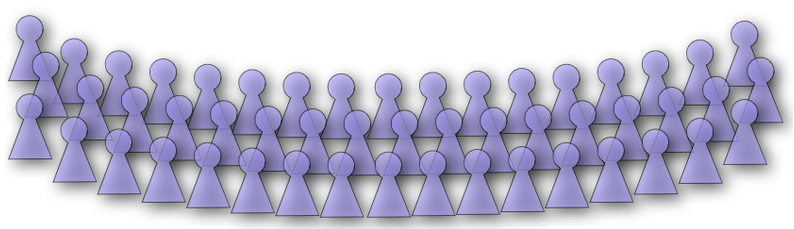




Exploration

- Presentation
  - Presents interpreted results
  - Tailored to fit a broad audience
  - Extensive scales, grids and legends
  - Static print without interactions
    (there are interactive Infographics by now)





Presentation

The Use of Graphics: Part II

- Exploration
  - Aims at gaining insights
  - Mainly personal use
  - Few scales and legends
  - Highly interactive and little persistent

- Diagnostics
  - is build upon standard plots
  - sometime a black box which is not well understood
  - often very specific to a statistical model or procedure
  - needs to link back to the raw data in order to improve the model

- An exploratory integration of diagnostic plots should deliver the best of two worlds

Where does R fit into this Taxonomy?

- Kinds of Graphics R supports
  - Diagnostics ✔
  - Presentation ✔
  - Exploration ??

- Diagnostics
  - almost all statistical procedures have at least some diagnostic plots
  - sometimes it is hard to link back to the original data or the model's parameters and settings

- Presentation
  - Many options and many packages as long as it does not get interactive

- Exploration
  - brush() and spin() in the old S-Plus days
  - iplots package

Where does R Graphics come from …?

- The core of R-Graphics (which is actually more S-Graphics) is based on a pen-on-paper model as was state-of-the-art at that time – and maybe older than most of the audience
- The explicit use of graphics devices makes porting easier but limits features somewhat to the lowest common denominator
- R's package system gives us an enormous flexibility to extend and build upon existing components – but only within R's technical limits, i.e., little interaction and single thread
- … but modern visualization utilizes
  - interactions, and
  - animations
- The easiest way out is to run graphics in parallel and talk to R

R-Graphics Packages

- General purpose
  - **graphics** Standard graphics grown over decades
  - **lattice** based on **grid** R incarnation of the Trellis library known from S+
  - **ggplot2** High quality presentation graphics
  - **iplots** High interaction graphics compatible to graphics

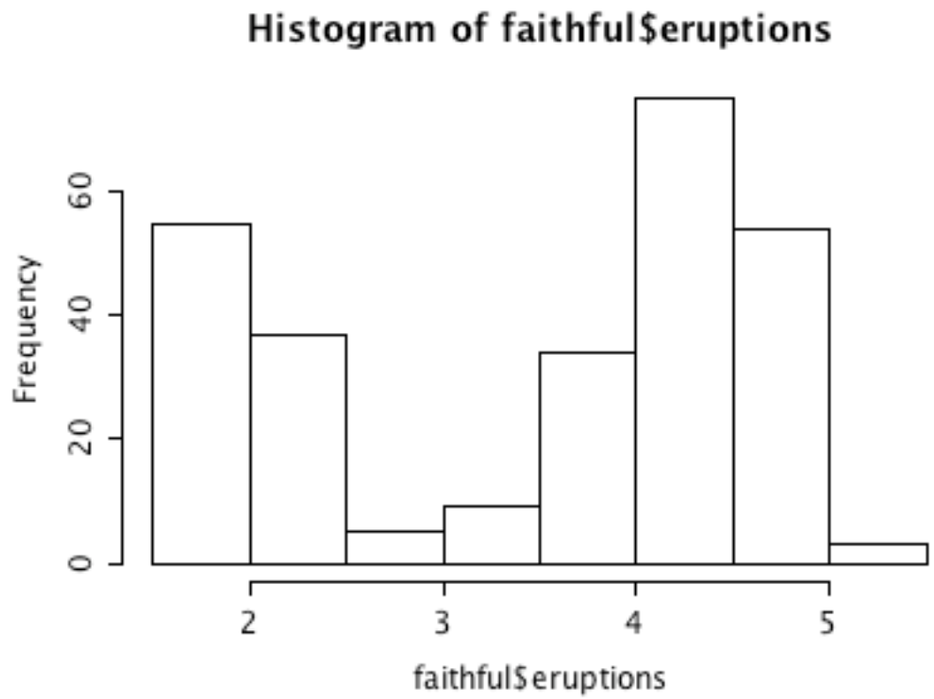
- Special domains
  - **party** “finally decent trees in R”
  - **vcd** Adds all sorts of visualizations for categorical data
  - **...** … *you name it*

- Graphics software which communicates with R
  - ggobi via **rggobi** “remote control” for ggobi in R
  - **KLIMT** pioneer in using R as a slave
  - **Mondrian** (to come) future release planned to control Mondrian from within R

How do I draw a Histogram in R?

- You use package: base

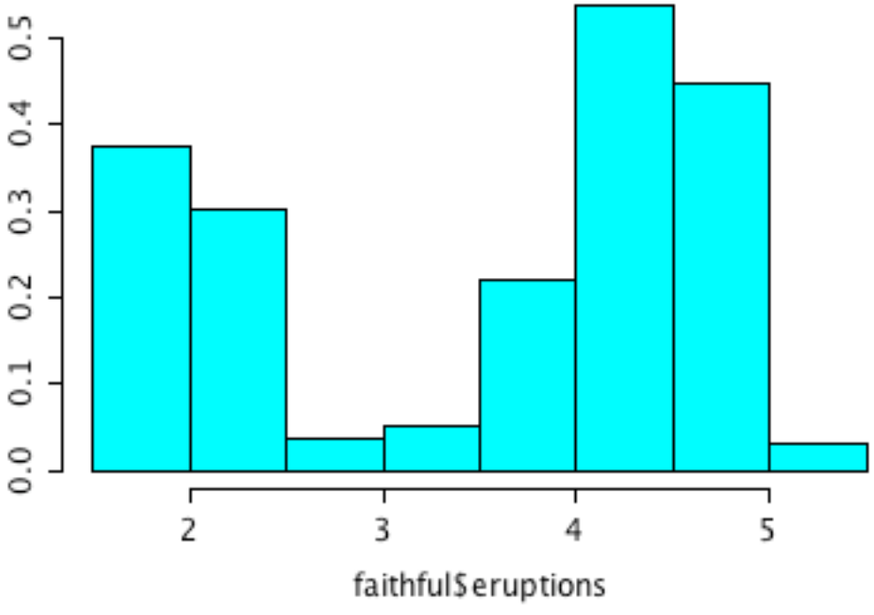
```
hist(faithful$eruptions)
```



How do I draw a Histogram in R?

- You use package: base , MASS

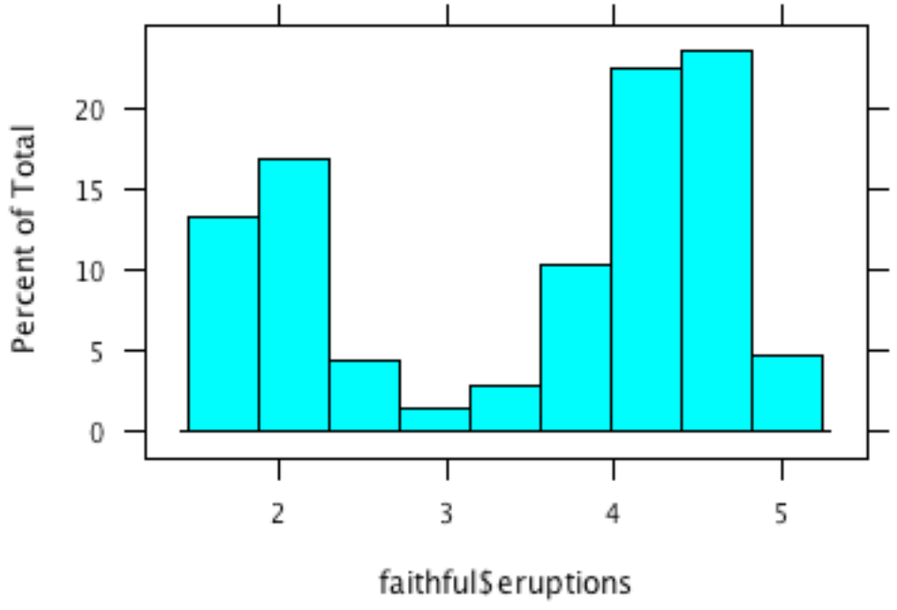
```
truehist(faithful$eruptions)
```



How do I draw a Histogram in R?

- You use package: base , MASS , lattice

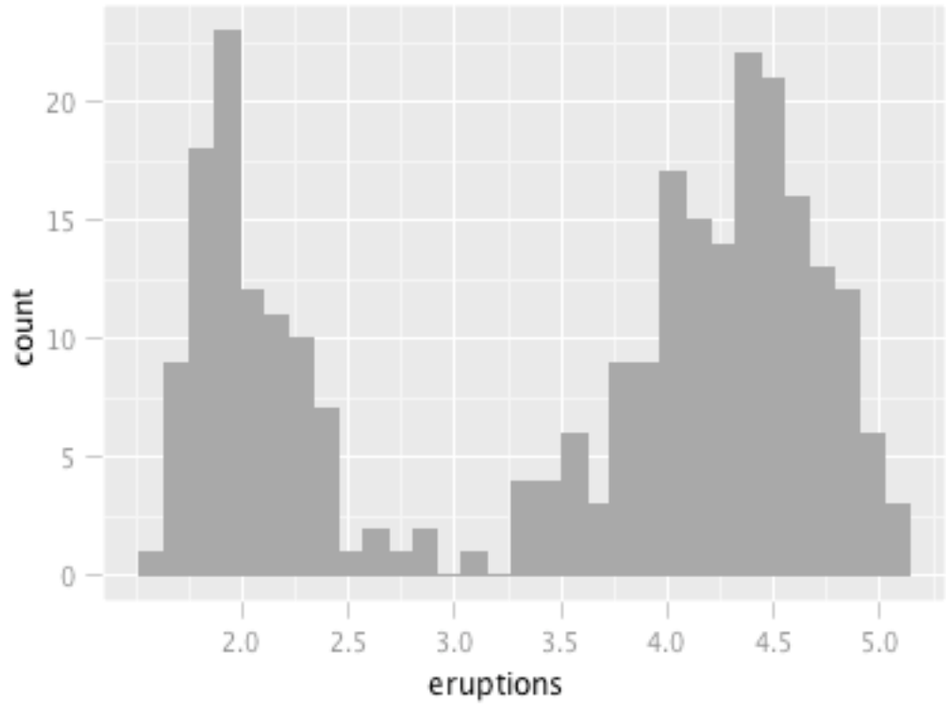
```
histogram(faithful$eruptions)
```



How do I draw a Histogram in R?

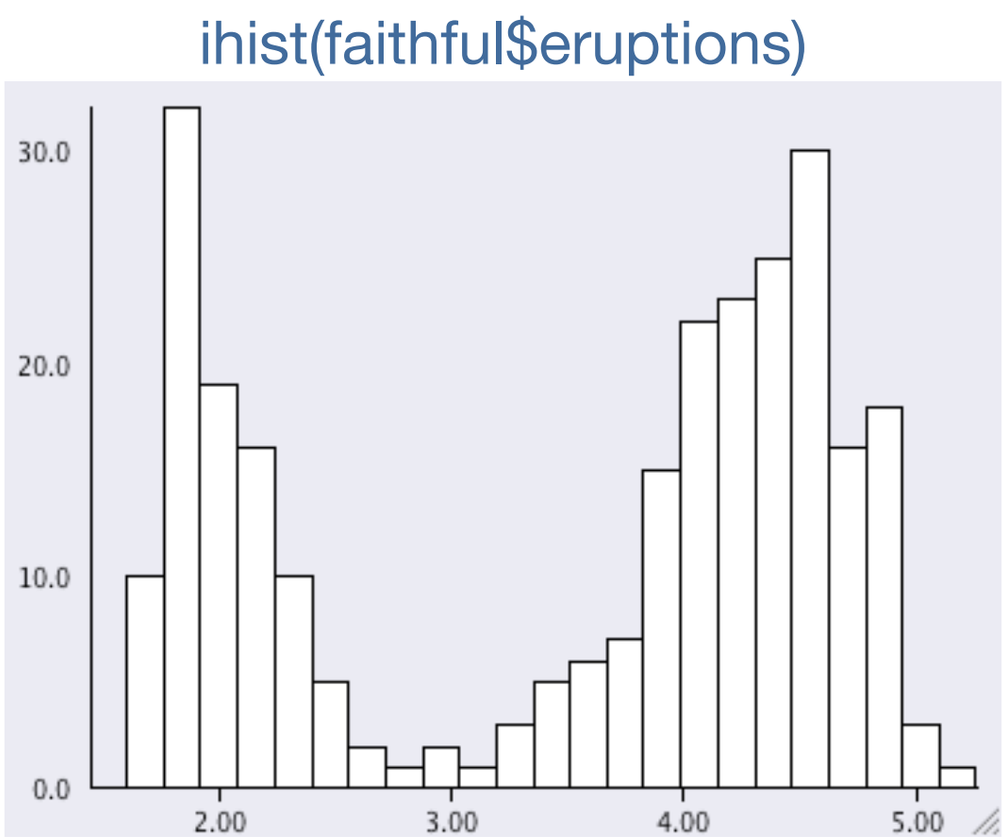
- You use package: base , MASS , lattice , ggplot2

qplot(eruptions, data = faithful, geom="histogram")



How do I draw a Histogram in R?

- You use package: base , MASS , lattice , ggplot2 , iplots , …



- All solutions have their specific pros and cons, but do we still speak the same language?

iPlots eXtreme

- What we didn't like about iplots so far …
  - JAVA graphics is just too slow when it comes to really large data (despite all the promises of SUN to support OpenGL within JAVA)
  - We kept a copy of the data both in R as well as in the JAVA VM
  - The user interface somehow strayed towards featurism and got clunky

- What iplots eXtreme (will) offer(s)
  - Snappy interactions far beyond 1 million items in all plots (the power of your graphics chip will be the bottleneck)
  - No more copying of data, only references are used
  - Cleaned up user interface

- iplots eXreme Goodies
  - Extensibility (custom objects that are really interactive)
  - Can be used as an ordinary (very fast) device
  - Offers built in support for interactive visualization of statistical models

Where do we go from here?

DISCLAIMER: this is by no means objective, but more a personal wish list

- "clean up"
  although the R-foundation never claimed the authority to do so, I see is a certain need to consolidate packages – not only in graphics
- "interaction"
  introducing a new standard device which allows for interactions (and more) will open R graphics to an even broader audience
- "graphics for exploratory data analysis"
  with ggplot2 we already have a great package for presentation graphics but exploratory graphics should be more "handy"
- "graphics for advanced model exploration"
  by now every statistical procedure in R has some diagnostics plot, but most of them fail to actually visualize the model

Summary

- We should always take John Tukey's "*There is no excuse for failing to plot and look*" to heart
- "*A picture is worth a thousand words*"
  is still (mostly) true, but as statisticians we should read it more like
  "*A full graphical analysis involves drawing a thousand pictures*"
- Following only a few guidelines, we can make sure that we create sensible (non-standard) plots that transport the right message
- Exploration graphics and diagnostic graphics should more and more become one as they serve the same goal – data analysis
- R offers great extensibility of graphics packages but all solutions that offer interactions with the plots are still patchwork
- iplots eXtreme may be a good start into the right direction