Data Visualization and R

Theory and Implementation
Spring 2016

Ryan Womack(rwomack@rutgers.edu)
Data Librarian, Rutgers University

March 31, 2016

Introduction

What this workshop IS:

- Focuses on standard techniques of data visualization, the day-to-day power tools for understanding data
- Reviews various graphical techniques, from early to recent, from simple to advanced
- Discusses principles of good data presentation, and show the R implementation of many functions

Introduction

What this workshop is NOT:

- It is not about "infographics", the beautiful, heavily customized products of expert graphic designers
- Not about the cognitive science aspects of data perception [wish I knew more about this!]
- It is not a complete introduction to R, even though R is used
- It is not an introduction to other software beyond R, or the use of R with scripting languages to produce interactive graphics on the web
- It is not necessarily a balanced survey of all data visualization. In particular, it is light on graph networks, clustering, and trees [not my expertise]
- Very little mapping, too. [Done to death?]

Introduction

What you can hope to gain:

- Familiarity with the basic principles and history of data visualization, and recent developments
- Exposure to a wide-range of plotting techniques and R packages
- A sampling of interactive and big data methods
- Understanding of the power and potential of combining appropriate graphics techniques with data

Setup

- Workshop materials, including R scripts, supplemental images and data, are available for download from https://github.com/ryandata/DataViz
- The R script file contains working demonstrations of the concepts mentioned here.
- You will have to install any packages not already on your system.

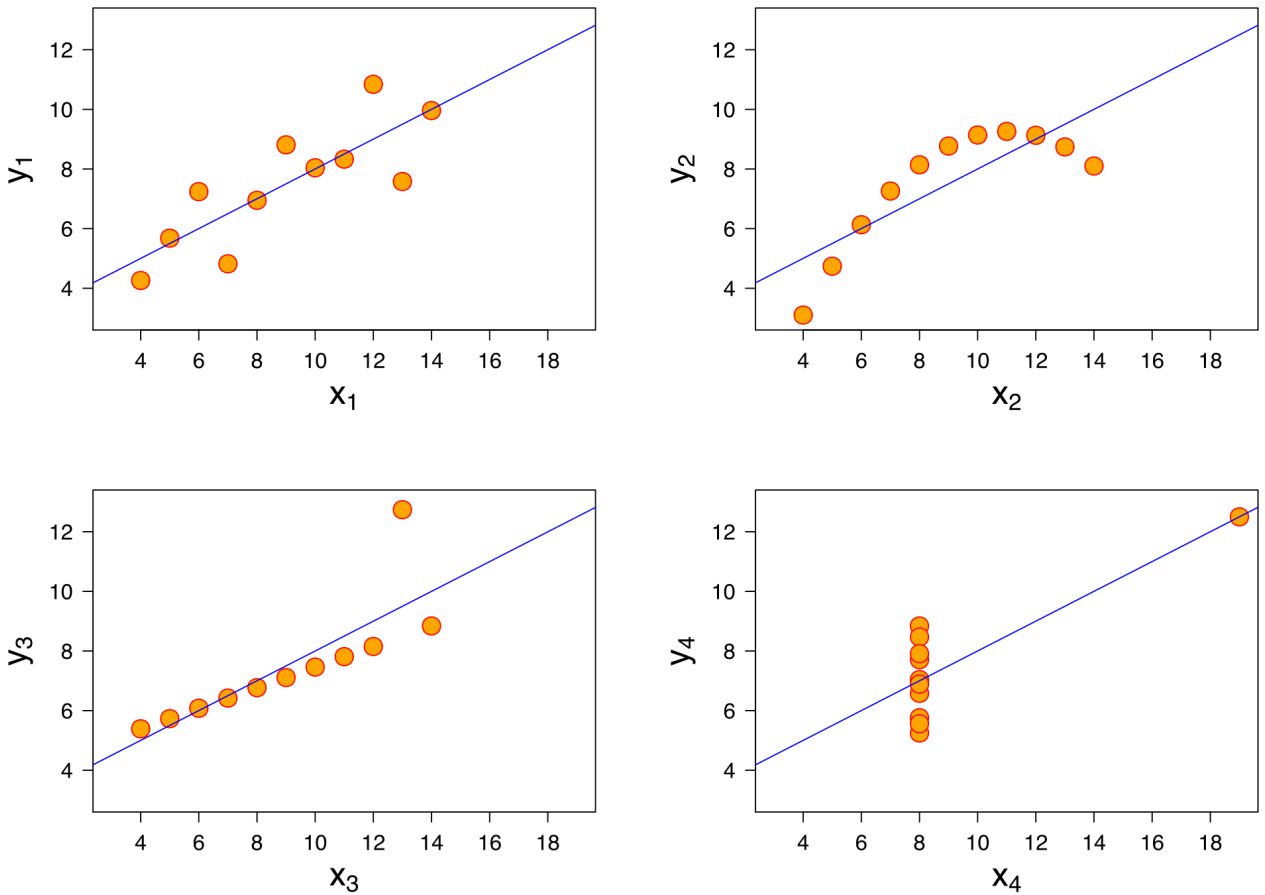
Why Data Visualization?

Data visualization can:

- provide clear understanding of patterns in data
- detect hidden structures in data
- condense information

Anscombe's Quartet

For example, see Anscombe's quartet (image source: http://commons.wikimedia.org/wiki/File:Anscombe%27s_quartet_3.svg):



Links to DataViz sites

Some examples of good data visualization (and fancy infographics) can be found at:

- Information Aesthetics
- Chart Porn
- Eagereyes
- DataVis.ca
- Visualizing.org
- VizWiz
- US Census Data Visualization Gallery

Bad Graphs

- Pie Charts are known to be problematic
- Clutter and other issues can ruin graphics

For more bad ideas, try:

- Junk Charts
- Ten Worst Graphs
- WTFviz

Pie Chart Examples

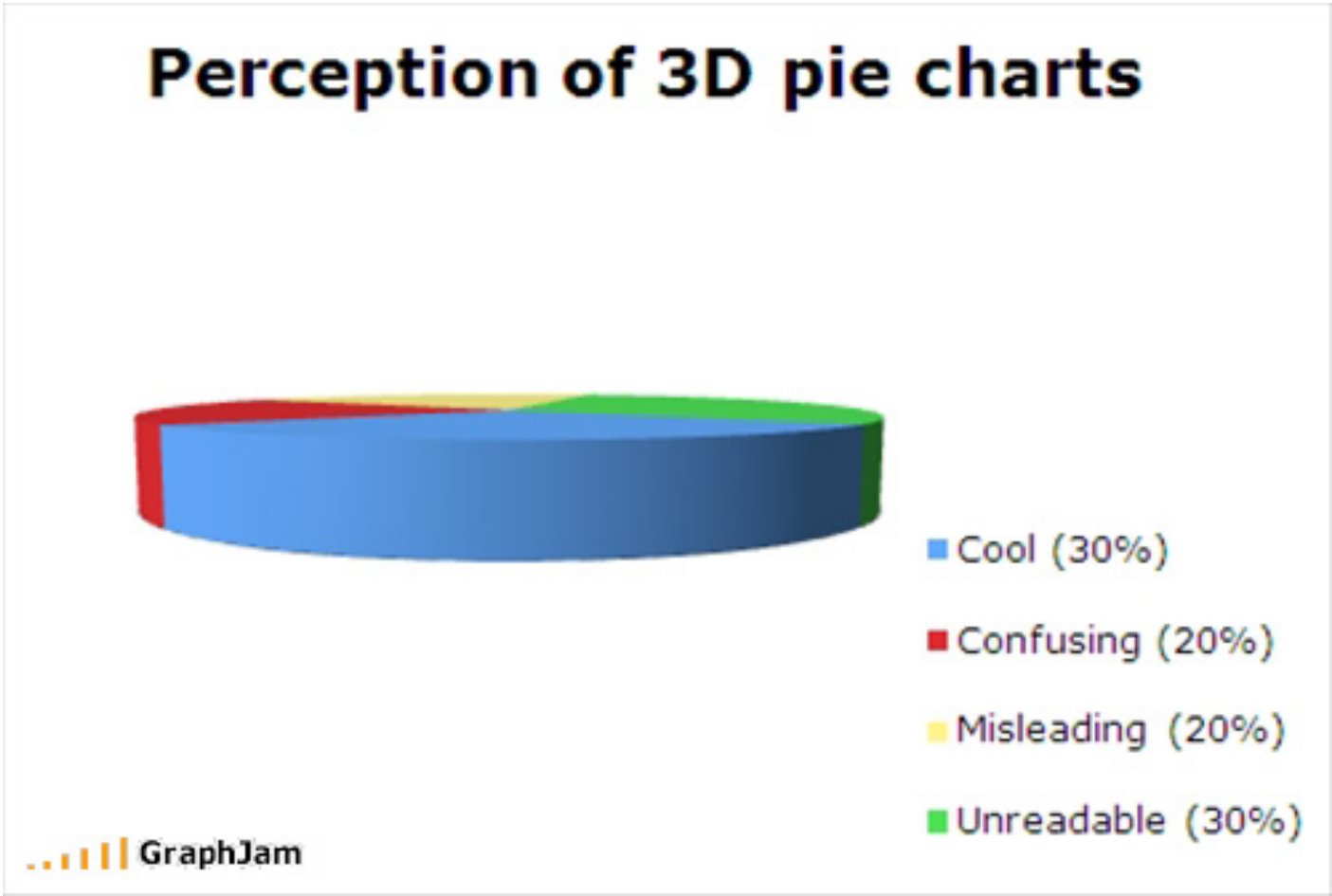


image source: http://peltiertech.com/WordPress/3d-pie-charts/

Pie Chart Examples

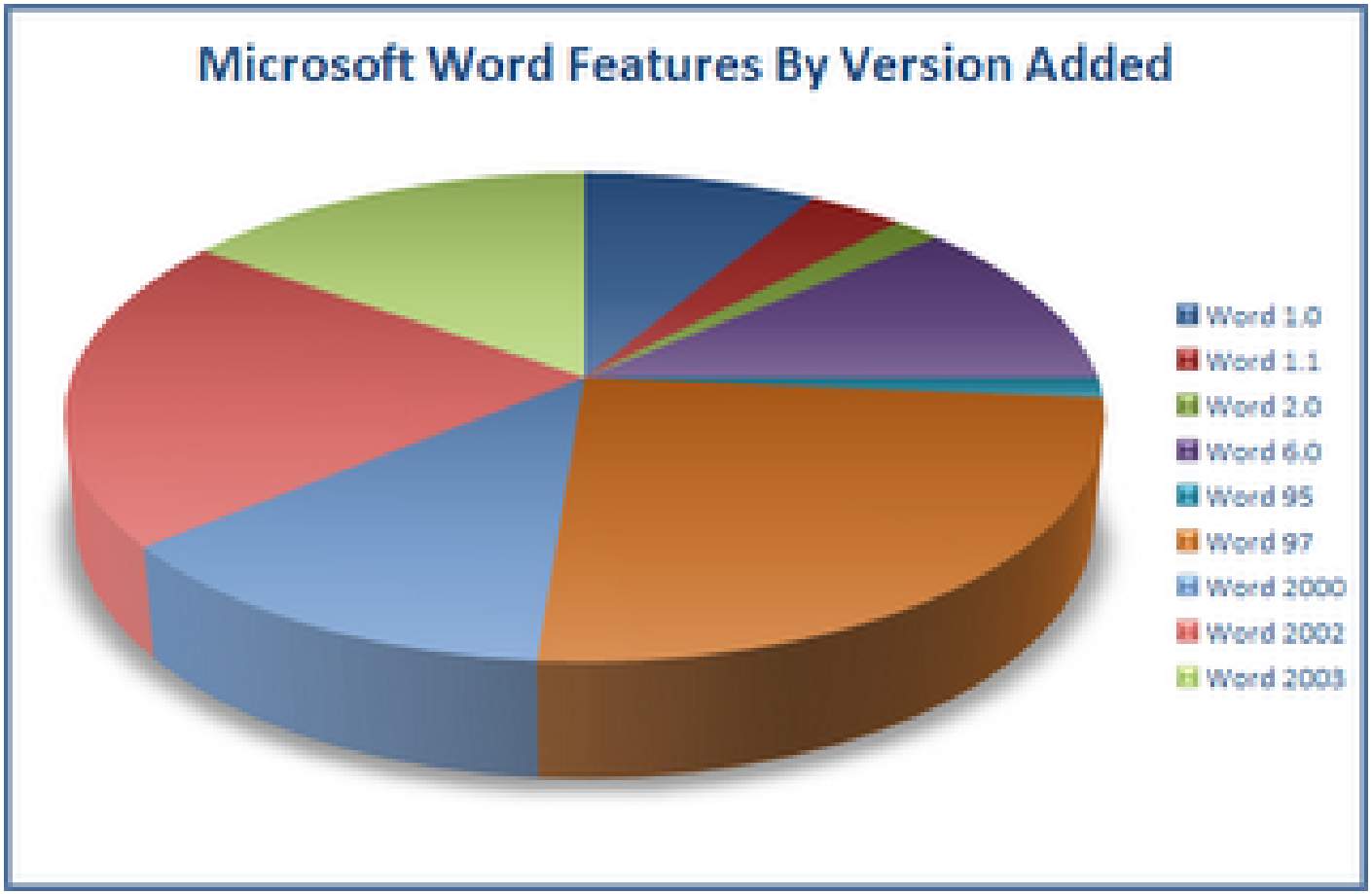


image source: http://ndevisual.wordpress.com/tag/uses-of-pie-charts/

Pie Chart Examples

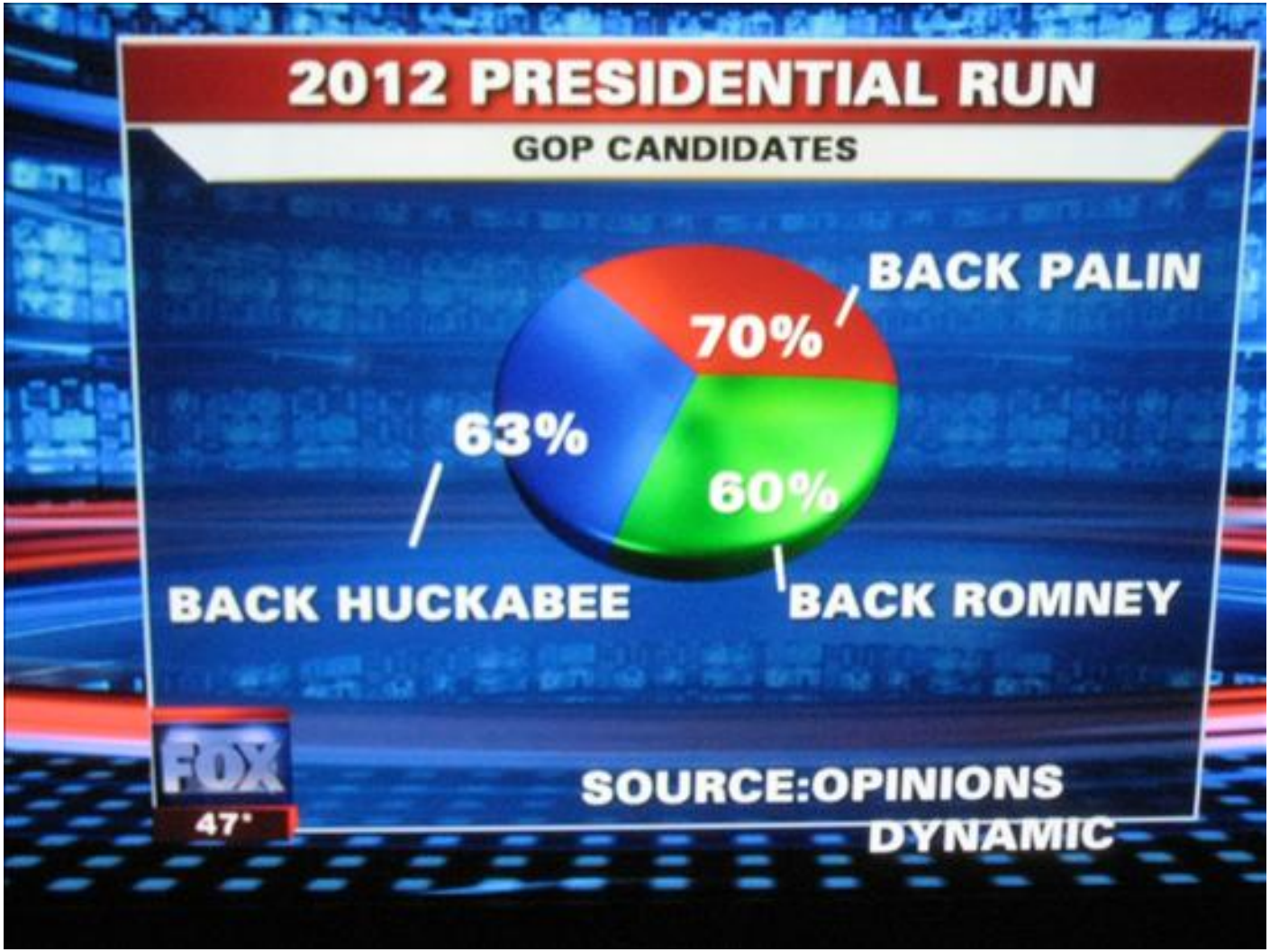


image source: http://www.nbcchicago.com/news/local/FOX-News-Chart-Fails-Math-73711092.html

Pie Chart Examples

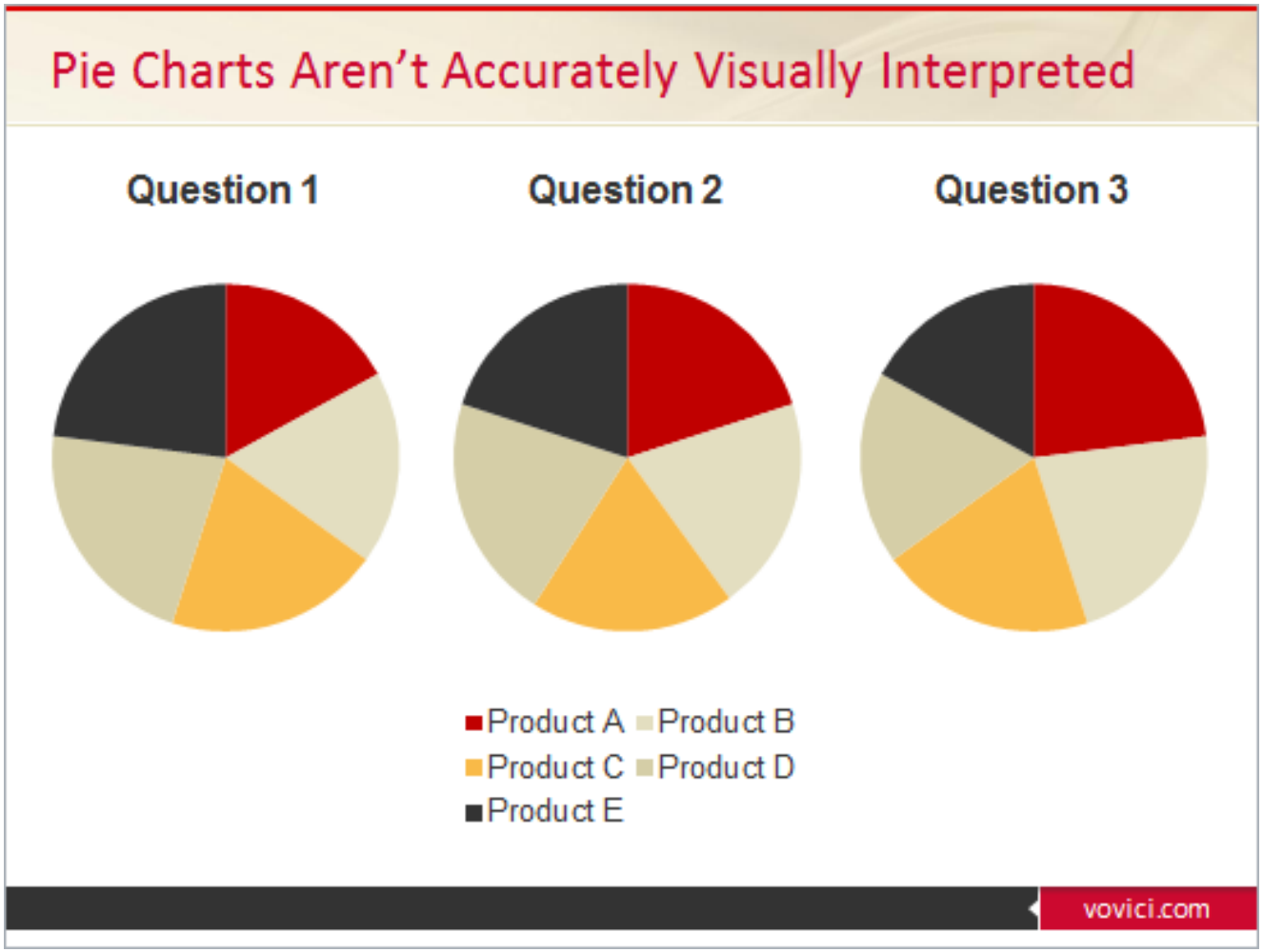


image source: http://tips.vovici.com/content/111031_swb

Pie Chart Examples

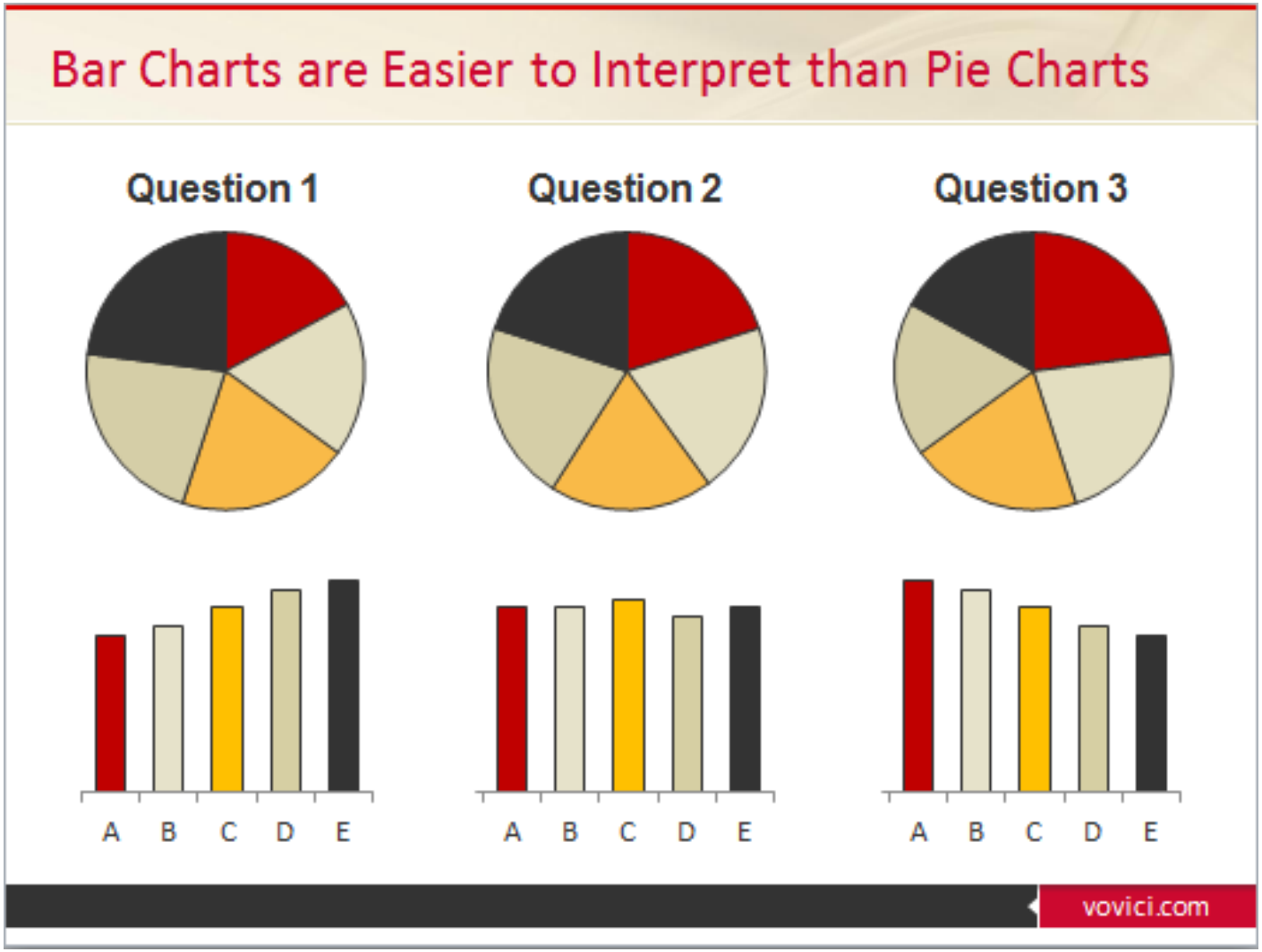


image source: http://tips.vovici.com/content/111031_swb

Clutter Example

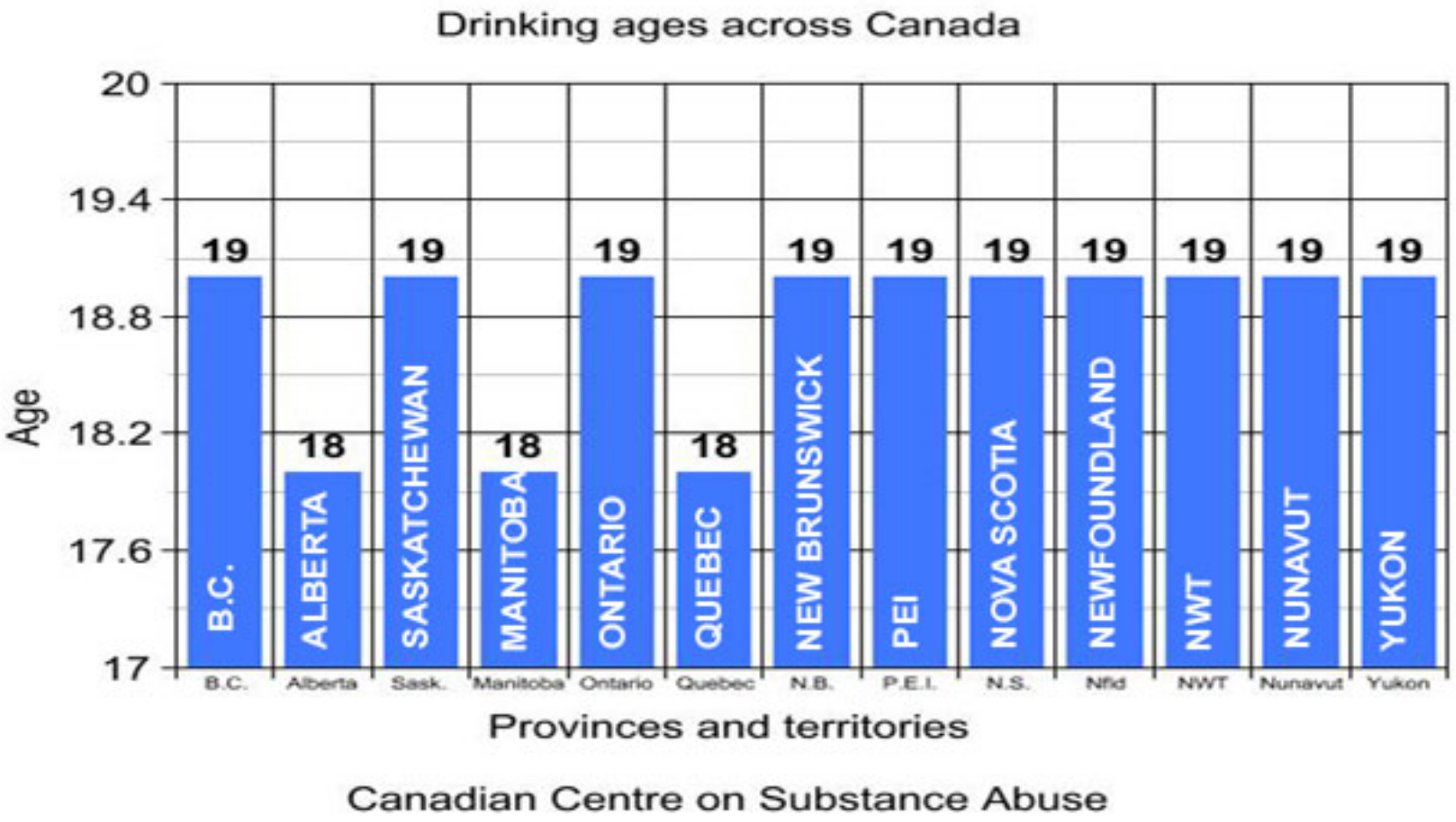
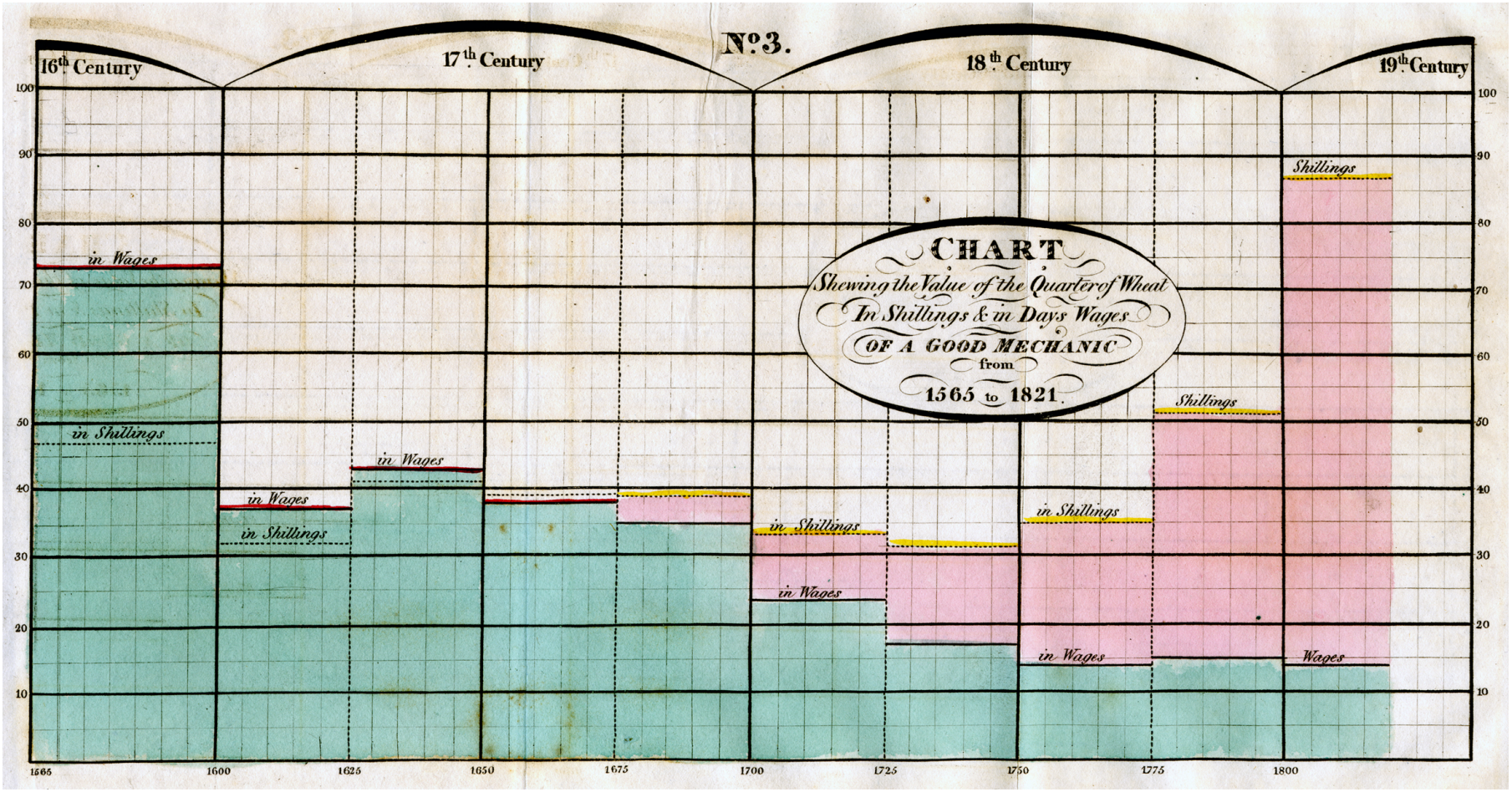


image source: http://junkcharts.typepad.com/junk_charts/2013/03/which-software-is-responsible-for-this.html

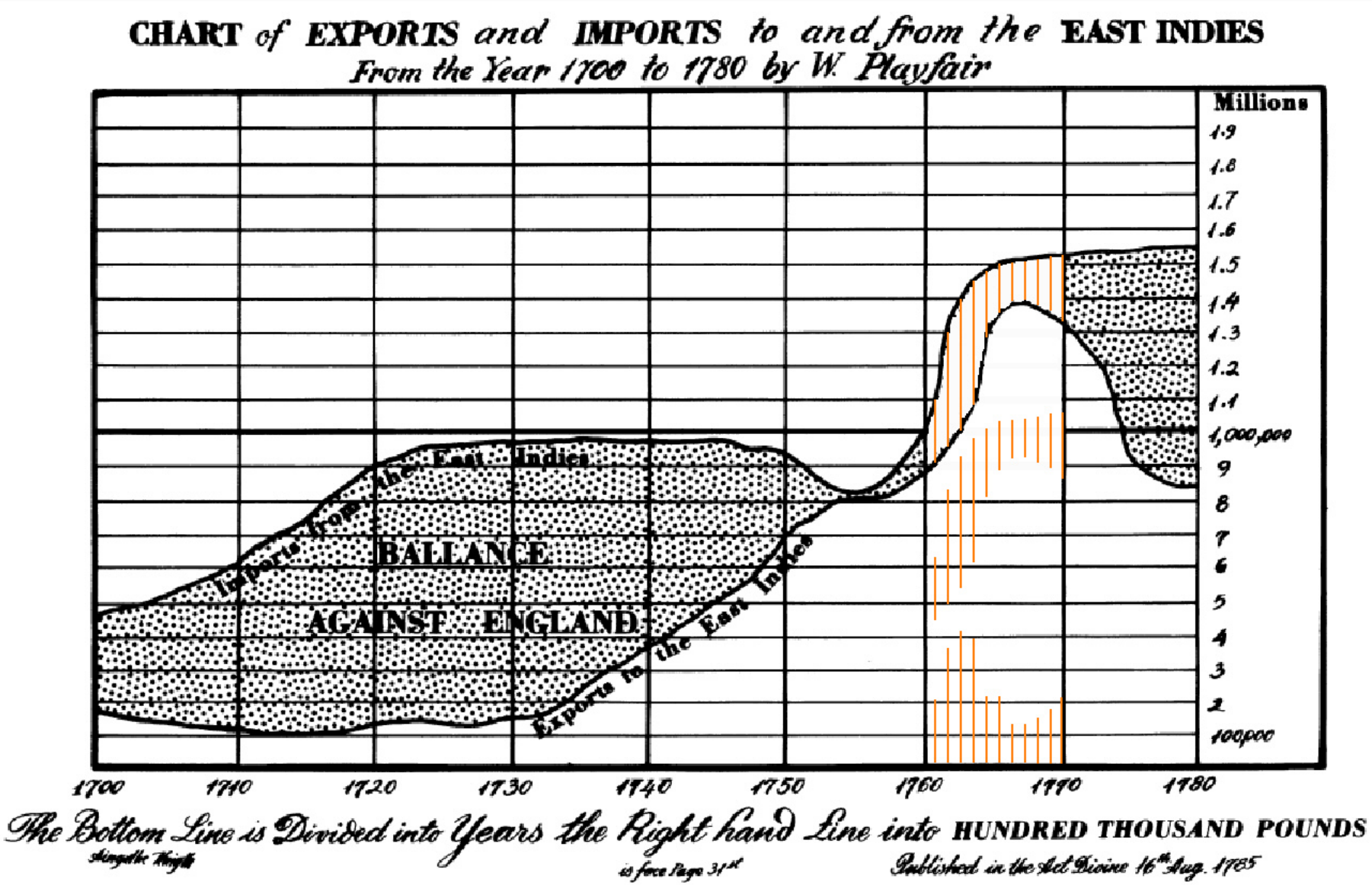
Playfair

- Astronomical observations, charts, and maps led in graphical innovation prior to 1800.
- William Playfair is the pioneer of the line chart, bar chart, time series plots, and pie chart.
- Playfair, W. (1786). *Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the Whole of the Eighteenth Century*,
- Playfair, W. (1801). *Statistical Breviary*.
- Both republished in *The Commercial and Political Atlas and Statistical Breviary*, 2005, Cambridge University Press.

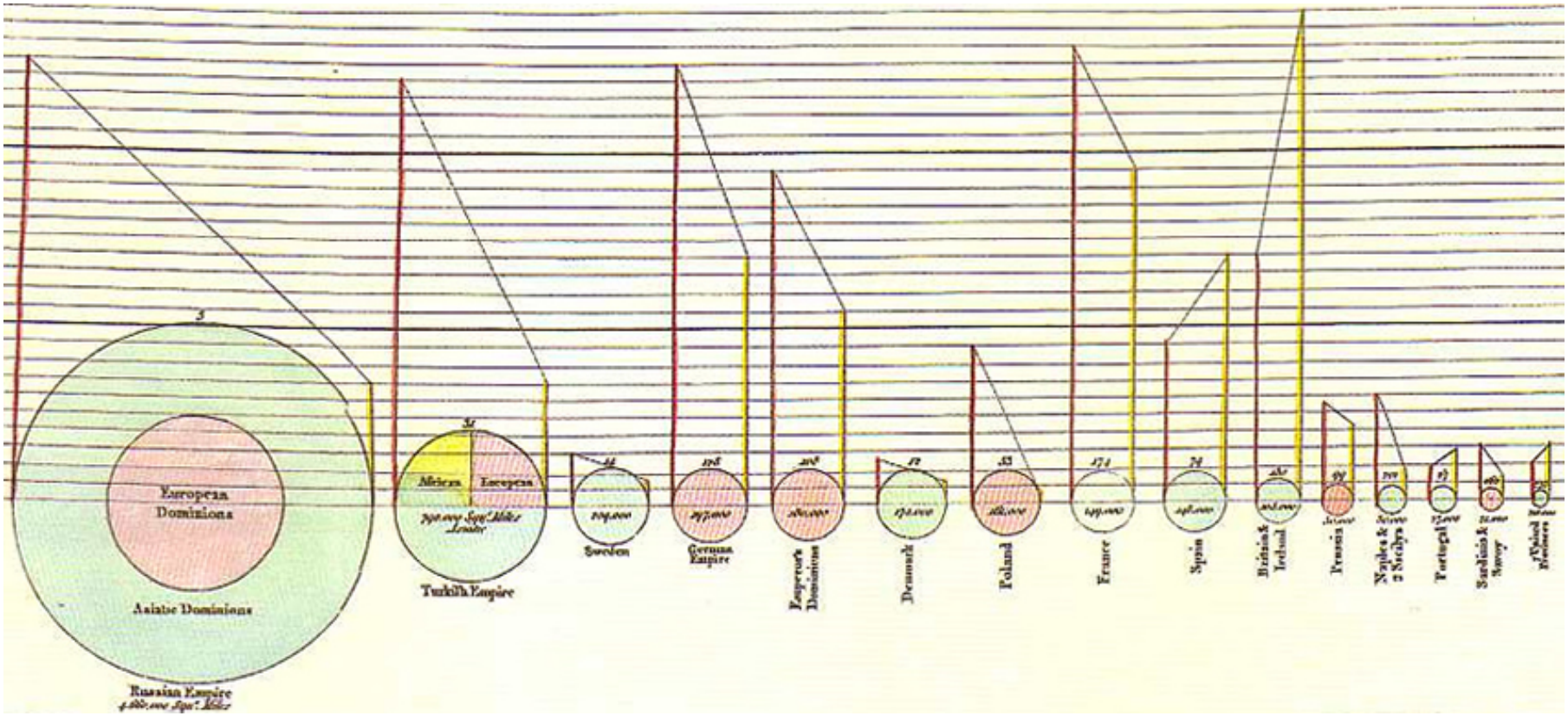
Playfair Examples



Playfair Examples



Playfair Examples



Piketty and Playfair?

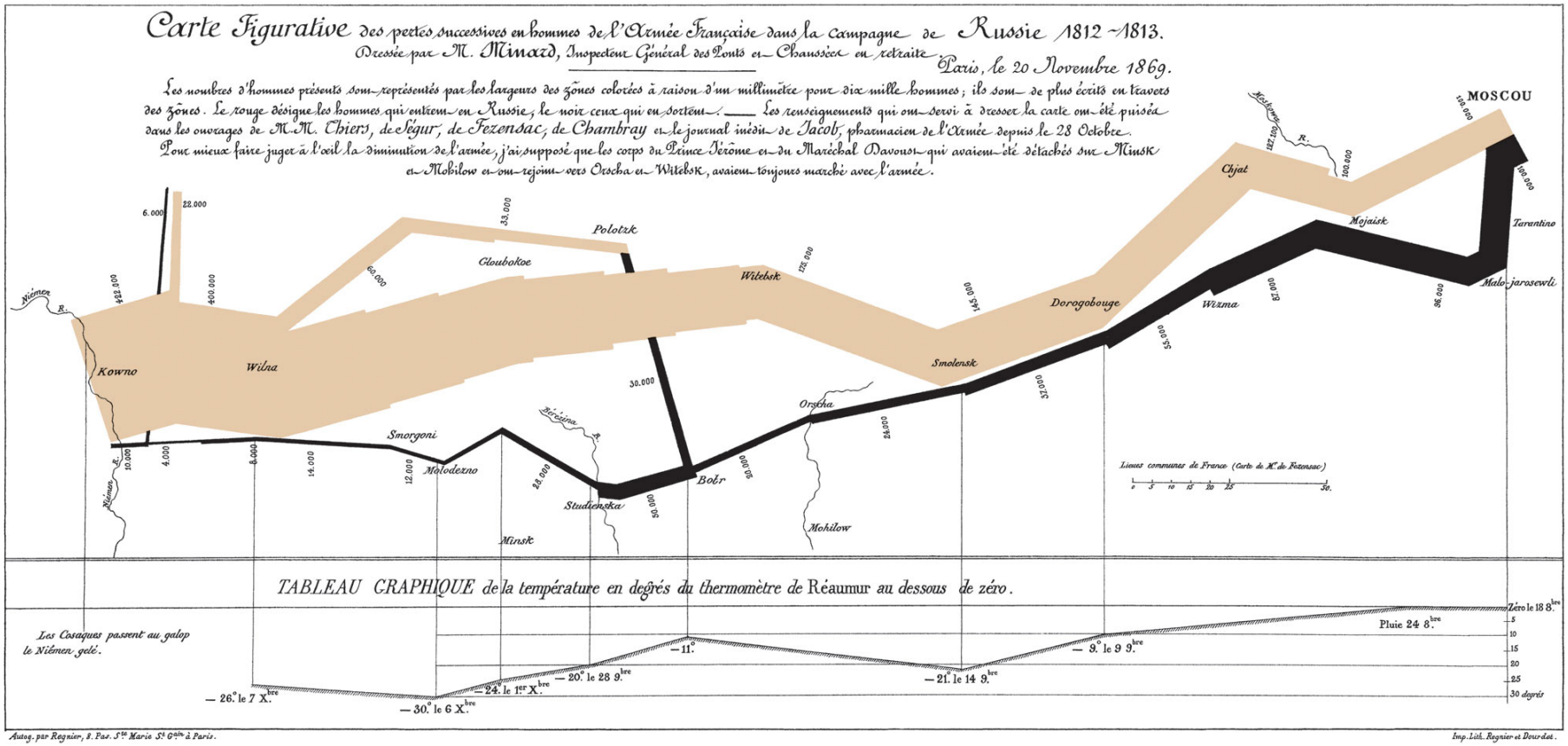
- Significance magazine suggests that Piketty's Capital garnered attention partly because of its careful use of figures in a style similar to Playfair's.
- But also that his charts could use some polishing!

Minard

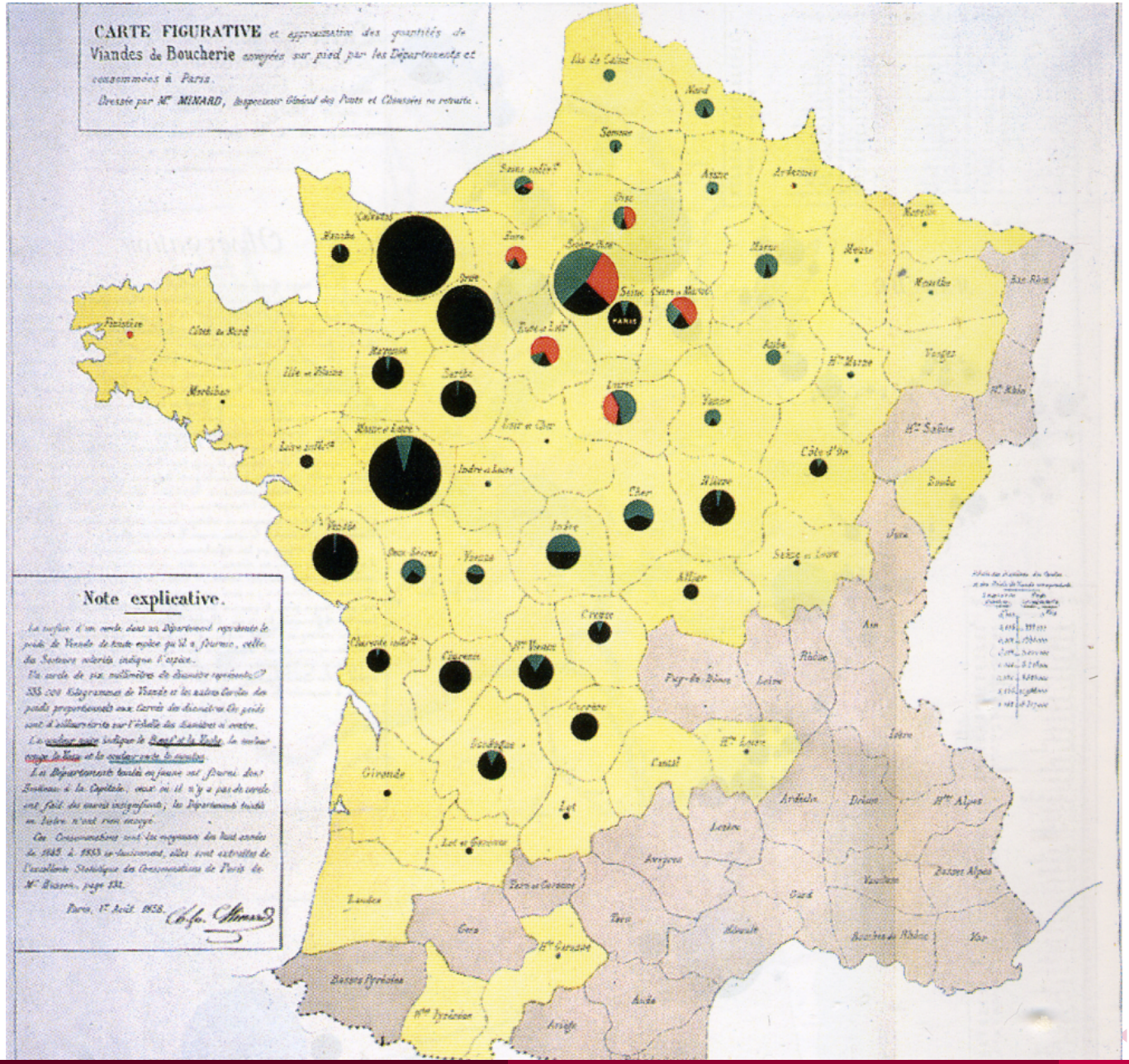
Charles Joseph Minard was the next influential data graphic creator after Playfair.

- Minard's flow map of Napoleon's Russian campaign is celebrated by Tufte and others as one of the greatest information graphics.
- It embodies an ideal of highly compressed informative elements, presented with style
- Six variables: size, location in 2 dimensions, the direction of the army, temperature, date [and group]
- However, this is a one-off design that crosses into Infographics, but it can be reproduced in R and other software.

Minard Examples



Minard Examples



Fisher and Tukey

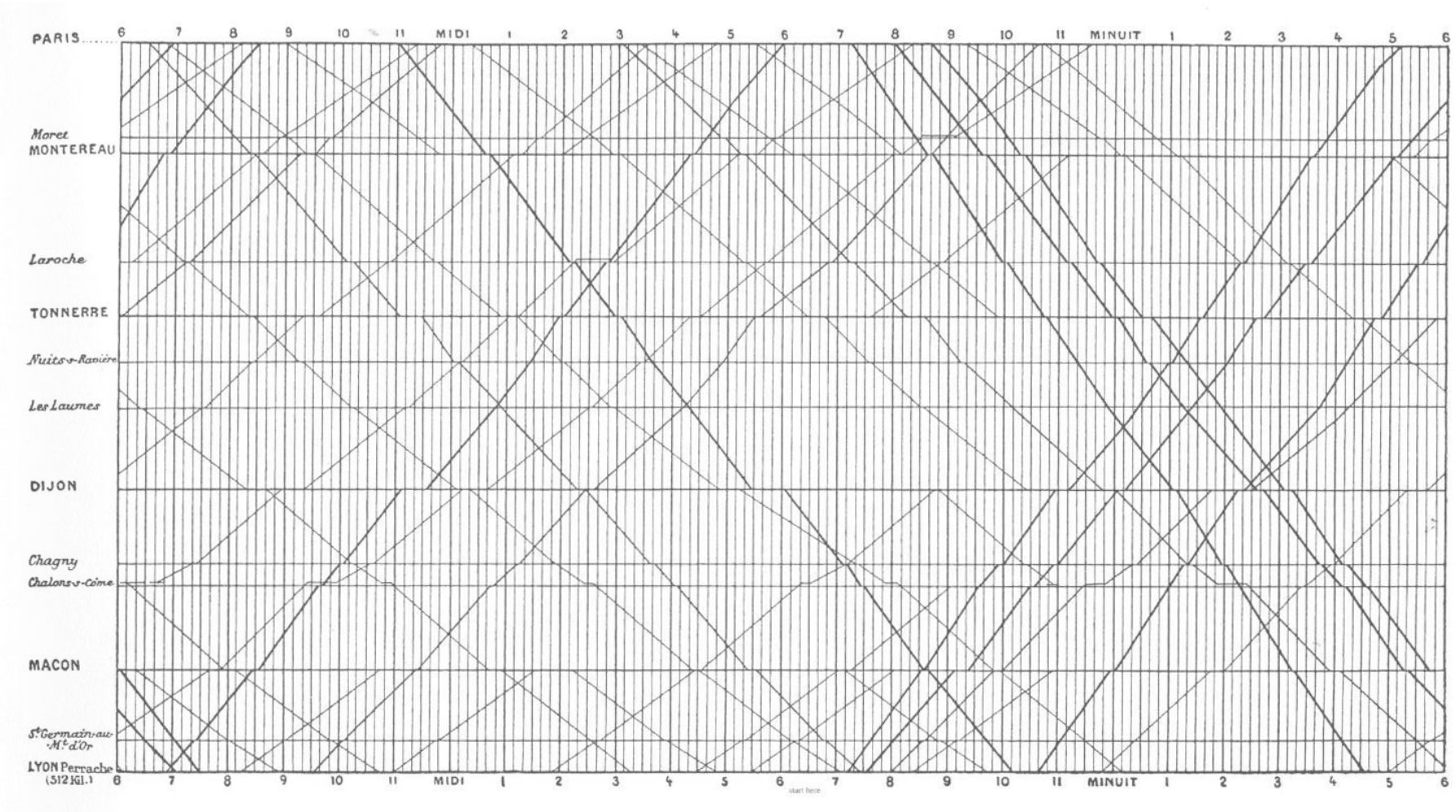
- Statisticians such as Ronald Fisher and John Tukey continued to advance graphical methods for the analysis of data.
- Fisher emphasized plotting the data to understand relationships.
- Tukey's *Exploratory Data Analysis* emphasized the use of graphics to understand the data during analysis, rather than the final presentation to an outside audience.
- Tukey created the box and whiskers plot and the stem and leaf plot.

Tufte

Edward R. Tufte's series of books, beginning with *The Visual Display of Quantitative Information*, have become the most widely know works on data visualization.

- There is considerable overlap between the various publications
- Tufte's ideal is highly compressed, elegant, and informative data, as expressed in dense printed graphics
- Tufte sometimes emphasizes beauty and design to the detriment of simplicity and clarity [e.g., train schedules]
- "Graphical elegance is often found in simplicity of design and complexity of data."
- "Beautiful graphics do not traffic with the trivial."

Train Schedule from Marey



Tufte's principles

Tufte has developed and popularized numerous principles and terminology:

- **Graphics reveal data** - show the data without distorting it - "above all else show the data"
- **Small multiple** - understanding one slice makes understanding others easier
- **Lie factor** - effect shown/effect in reality
- **Graphical Integrity** - no lies, let data vary, not design
- **Data density** - maximize data/ink ratio
- **Sparklines** - seems they haven't caught on
- **chartjunk** - self-explanatory
- **Powerpoint** is responsible for most of the world's sorrows [*The Cognitive Style of Powerpoint*]

Lie Factor

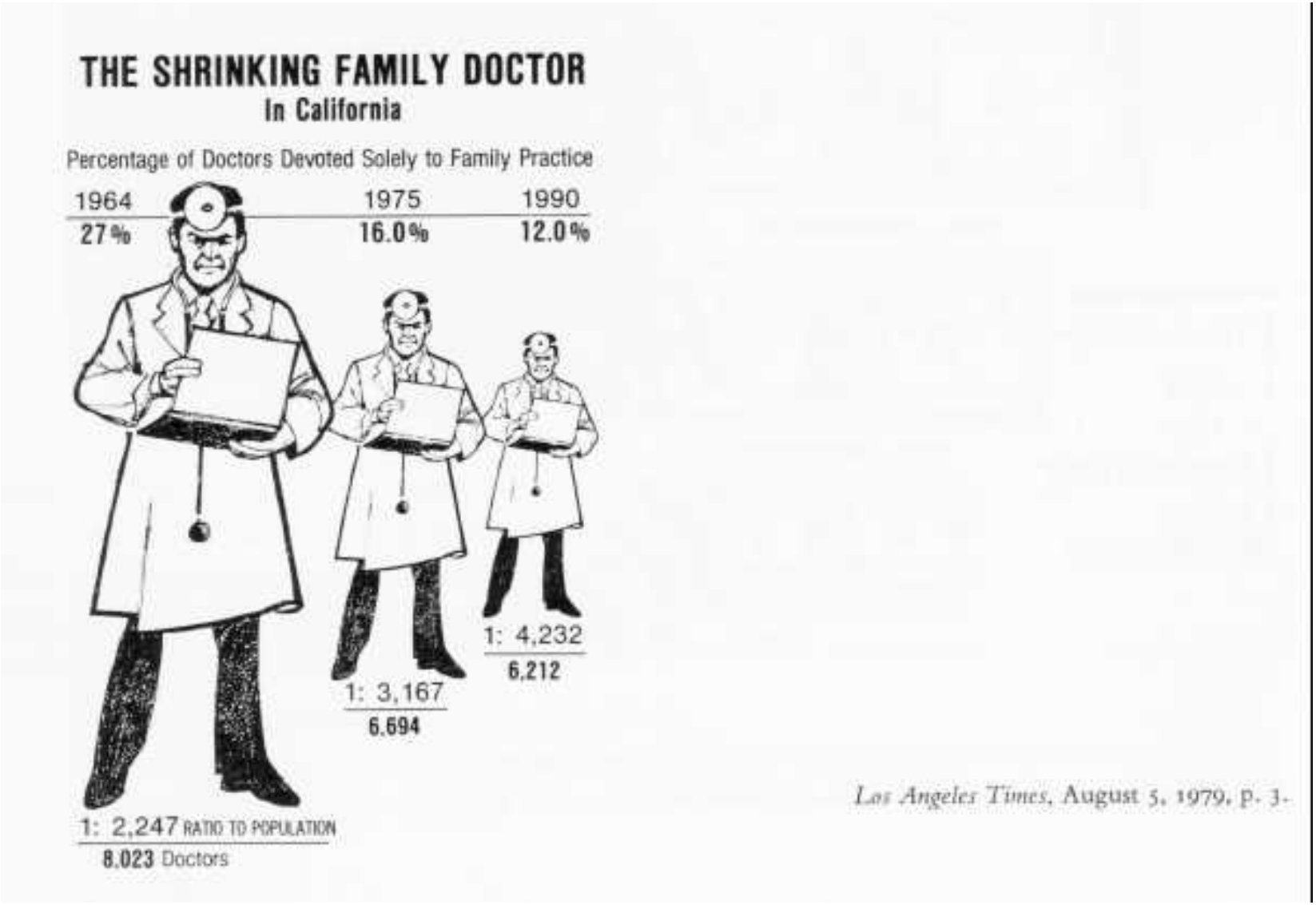
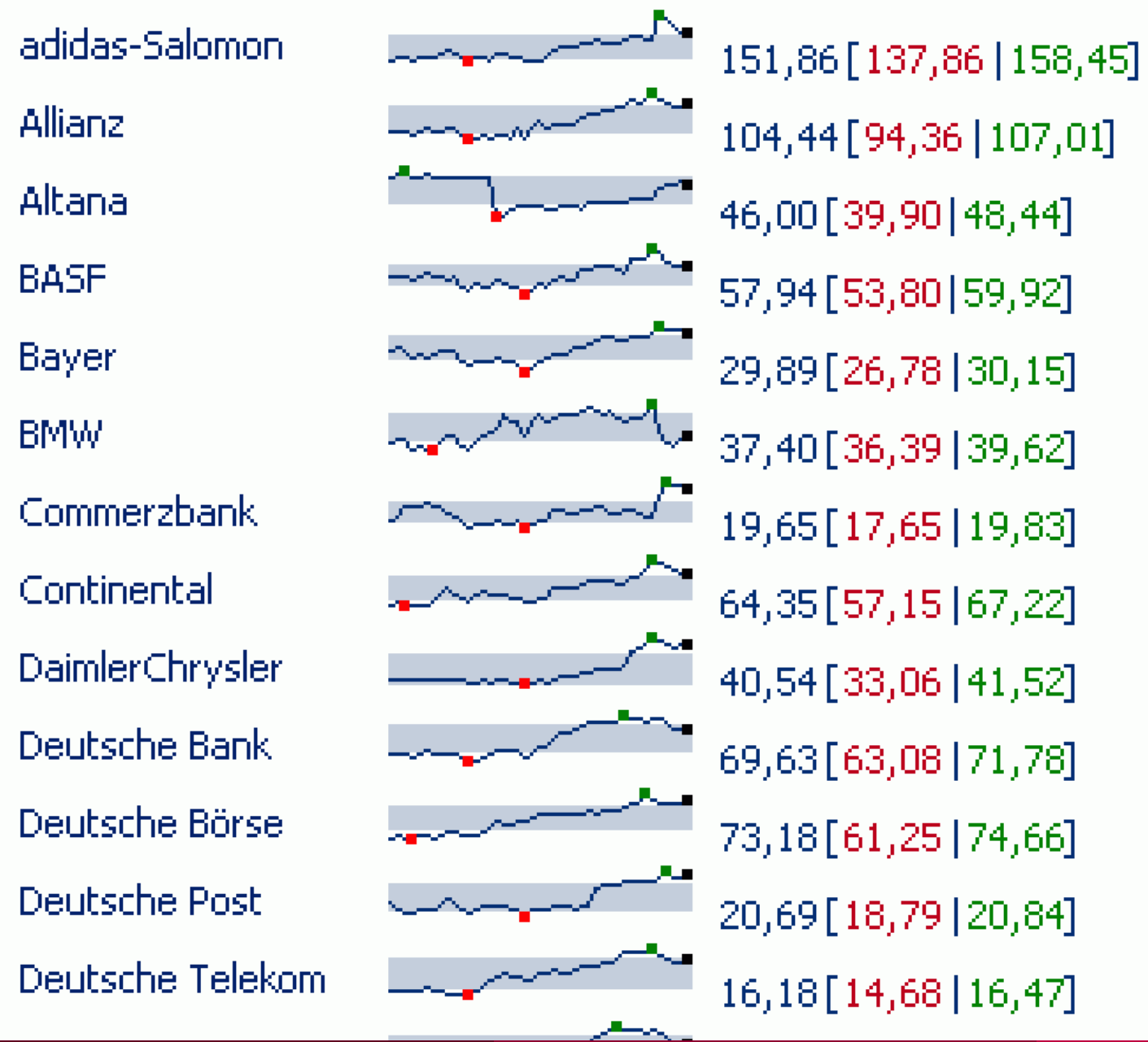


image source: http://www.datavis.ca/gallery/lie-factor.php

Tufte sparklines



Tufte counterexample - Minard revisited

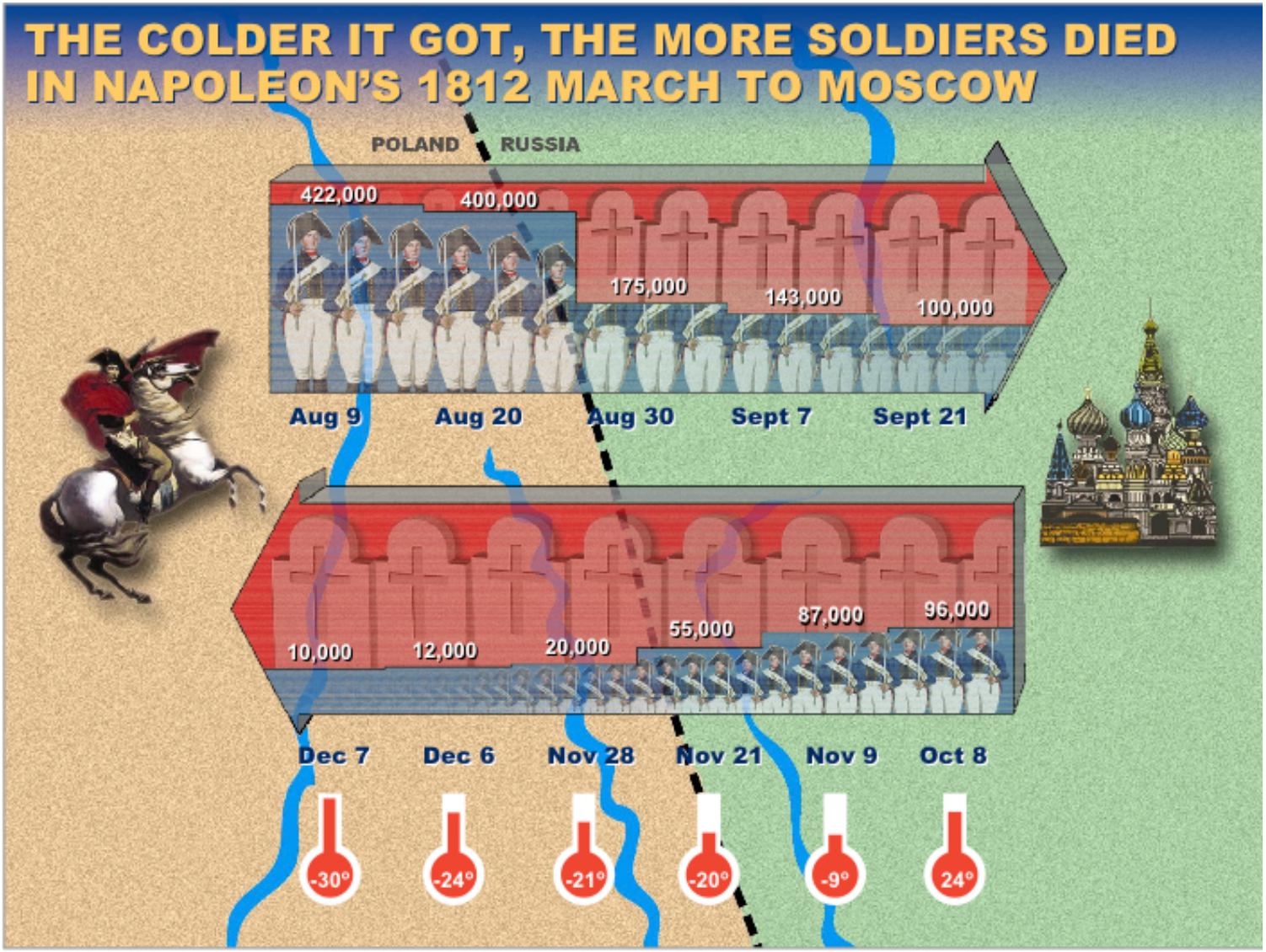


image source: Re-visions of Minard at http://www.datavis.ca/gallery/re-minard.php

Back to Pies

Why is the pie chart bad?

- Low data density
- Failure to order numbers along a visual dimension
- Perception difficulty in judging area
    - Stacked bar charts also pose perceptual problems

Or in the terms of Gary Klass:

- Data Ambiguity
- Data Distortion
- Data Distraction

A Good Pie Chart?

image source:
http://sciencegeekgirl.wordpress.com/2008/11/14/a-true-pie-chart/

Cleveland

- William Cleveland's *Elements of Graphing Data* and *Visualizing Data* pioneered systematic considerations of data legibility
- Cleveland is particularly known for promoting the *dot plot* as a alternative to bars and pies.
- The dot plot provides clarity and easy comparison of data.
- Cleveland also pioneered Trellis graphics
- Trellis graphics emphasizes comparison of multiple panels of data
- The `lattice` package implements Trellis graphics in R
- See Cleveland.pdf for a summary of Cleveland's recommendations

TECHNIQUES

- logs, % change, residuals
- point graph [2d histogram], histogram, percentile graph [and with comparisons/reference line], box plot [Tukey]
- dot charts - best way to attach label to quantity, 2-way dot chart {multiway} grouped dot chart
- overlap is dealt with by jitter, distinguishable symbols {sunflower plots}, taking log or other transformation
- box plots for high multiples
- visually distinguish curves and points [this has gotten easy by now]

Cleveland - 3 or more variables

THREE OR MORE VARIABLES

- Framed-rectangle graphs
- scatterplot matrices
- interaction/brushing
- 3d wireframe or stereogram (points)

Cleveland - Perception

PERCEPTION

- pie v. dot chart
- distance and detection
- length in a stacked bar
- 45 degree banking [Tufte also recommends 1.5:1 horizontal to vertical ratio]
- strive for clarity

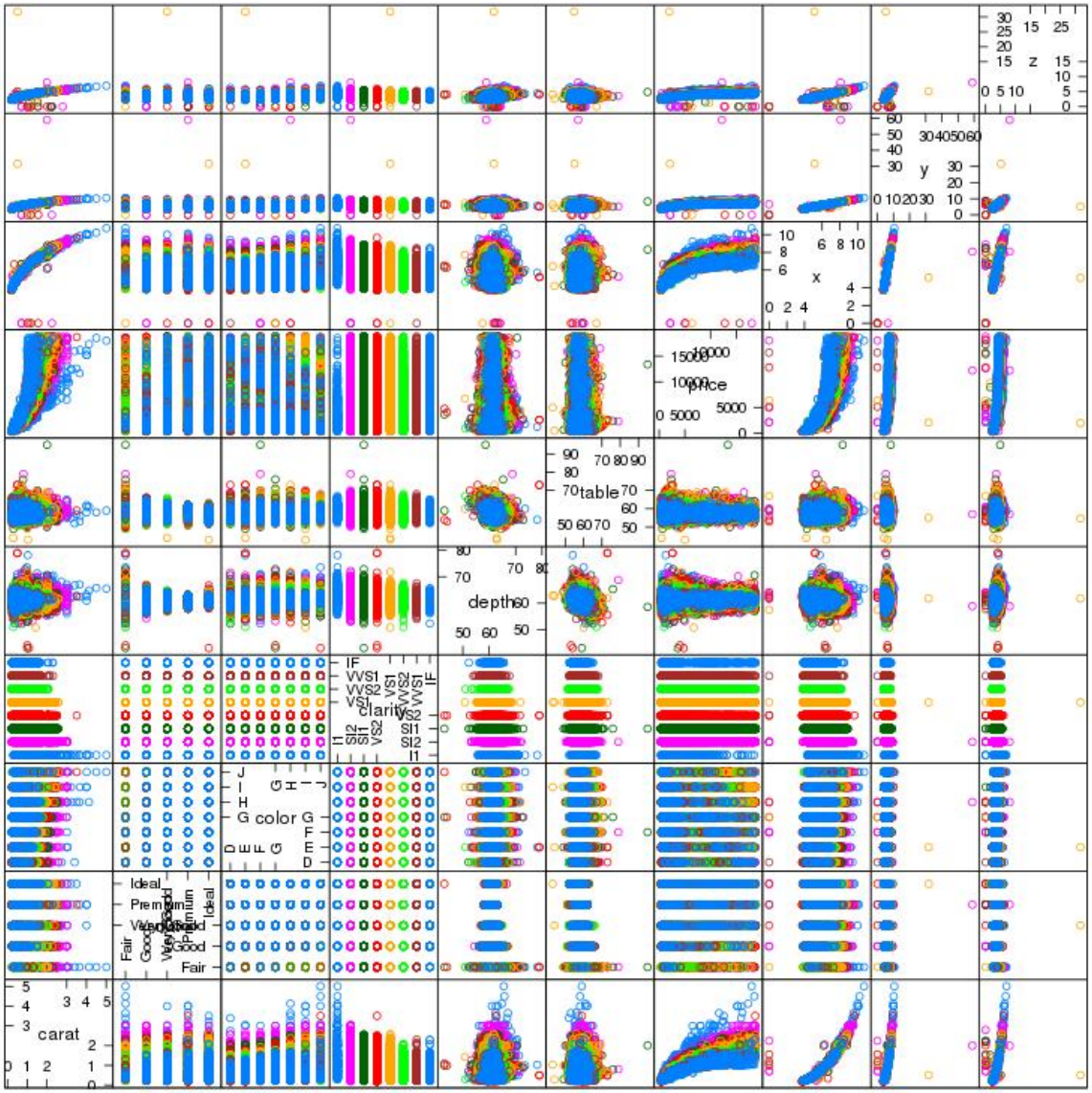
LATTICE

- The `lattice` package implements Trellis graphics in R
- `lattice` excels at comparative plotting
- uses similar syntax to base graphics, but with greater sorting and manipulative power

R basics

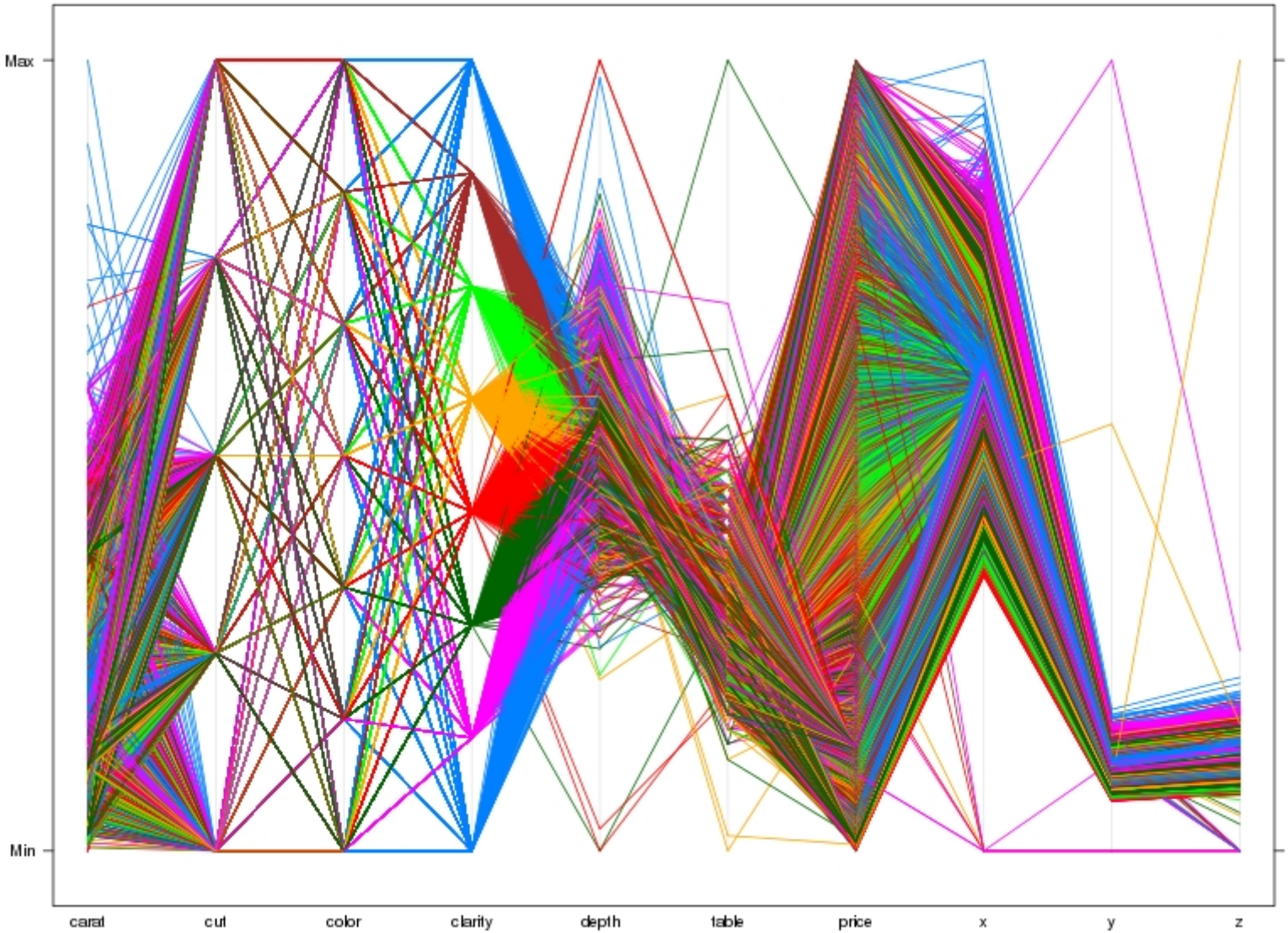
- install.packages, read.table, read.csv, library(foreign), library(xlsx)
- help via help.start(), ?, library(help="package"), vignette
- Cookbook for R and Quick-R are good jumping off points
- Guerilla Guide to R or R crib sheets or Rutgers R Libguide
- search Stack Overflow for R graphics to get specific coding tips on graphics
- R Graph Gallery has many examples of graphs with code, several adapted for this workshop
- Task Views or Crantastic can be use to discover pacakges. In addition to help and online documentation, packages are often described in articles published in places like the *Journal of Statistical Software*.
- This workshop will allow you to use the R language, but glosses over many details of structure and syntax to focus on the graphical elements
- You can always try R for cats if you get stuck!

SPLOM



Scatter Plot Matrix

Parallel Plot



The Grammar of Graphics

The Grammar of Graphics, by Leland Wilkinson, was extremely influential in thinking about graphics

- Grammar means ”rules for art and science”
- The Grammar of Graphics specifies rules both mathematical and aesthetic
- Earlier graph producers focused on aesthetics of static content
- Dynamic graphics and scientific visualization, by contrast, require sophisticated designs to enable brushing, drill-down, zooming, linking
- The Grammar of Graphics is easily adapted to this approach

Grammar of Graphics - pt I

- DATA - weighting, reshaping, counting, bootstrapping
- VARIABLES - transform, sort, log, ranking, residuals, quantiles
- ALGEBRA - nesting or blending data
- SCALES - nominal, ordinal, interval, ratio must be specified
- STATISTICS - static methods available to all graph types e.g, mean, sd, smoothing

Grammar of Graphics - pt II

- GEOMETRY - line, area, etc., along with modifiers like jitter and dodge
- COORDINATES - refers to the coordinate system of the graph (cartesian, polar, etc.)
- AESTHETICS - color, texture, size, position, etc. of the data points. Includes using color to classify.
- FACETS - subgroups, multiway tables
- GUIDES - legends, axes, color scales, keys

ggplot2

- `ggplot2` was developed by Hadley Wickham as an implementation of the Grammar of Graphics
- Relatively complete and powerful graphics package
- Can do many things, but not 3D
- See ggplot2 Help Docs and the ggplot2 book for complete descriptions
- Other short introductions to `ggplot2` are available, such as these from Sharp Statistics and inundata

A Miscellany of Visualizations

- The Cleveland dot chart
  - use to compare labeled quantities, ordered lists
- Kernel Density plot
  - visualize the distribution of data with more precision than a histogram
- Scatterplot Matrix
  - study relationships between all variable combinations

Visualizing Distributions of Data

- Box and Whiskers Plot
  - illustrate quantiles and outliers. There is also a Tufte version.
- Stem and Leaf Plot
  - see precise quantities associated with distribution of data
- Violin plot
  - Blends density information with box and whiskers style (in an artistic manner)
- Dot plot
  - plot distribution point-by-point
- Heat map
  - compare many magnitudes easily

Categorizing Data

Many techniques are available to automatically identify related data:

- A *tree* illustrates a categorical classification of the data based on its own characteristics
  - one implementation is the `party` package
- *Self-organizing maps* are a form of neural network that derives characteristics from the data and plots patterns
  - see `kohonen` and `som` packages
- *Clustering* of data can be accomplished by numerous algorithms
- `hclust` and `pvclust` are some methods described at Quick-R's Cluster Analysis page.
- Other graphs and network analysis tools are available [not explored here]

Visualizing Categorical Data

- The *mosaic plot* allows multiple categories to be displayed on the same graph, but can be complicated to interpret.
- The *spineplot* is a variant of the mosaic plot, plotting proportions in 2 dimensions.
- You can also do a timeline.

Maps and Glyphs

Maps are obviously an important and widespread way of presenting data.

- We examine a few examples of *choropleth* maps, in which shading indicates data levels

Glyphs present iconic representations of data elements as plotted points.

- Weather maps often use glyphs.
- A more dynamic example is here.
- As an R example, consider Chernoff faces and the `aplpack` package. Also, Smiley faces [and many more graph variants in this chapter].

3-D

- 3-D scatterplots
  - `cloud (lattice)`
- contour plots
  - to plot standardized levels of data
- wireframe plots
  - to present a 3-D surface representation of data
- `rgl` (a separate package containing several 3d plotting functions and animation)
- `mosaic3d` extends the mosaic paradigm to three dimensions

Animation

- Animation is an easy way to step through data over time
- or to provide comparisons of different views of data
- R makes animation easy with the `animation` package
- Just enclose a sequence of graphics in the animation command to generate interactive HTML (or GIF, SWF, LaTeX, Video).

Interactive DataViz - Principles

- Why aren't all of our graphs interactive?
- *Brushing* is used to select data points and track them through various analyses.
- Drilling down, zooming, and subsetting are also interactive techniques.
- Data displays can be linked so that a selection in one panel modifies the output displayed in another panel.
- Interactivity is especially useful for data exploration, studying multidimensional relationships.

Linked Data Panels and Vis packages

In many contexts, visualizing the relationships between data elements is made easier by viewing related data panels simultaneously.

- One example of this occurs in time series data with decomposition into trend, seasonal, and random components
- The tableplot (`tabplot` package) implements another linked data view across all variables.
- googleVis and other "Vis" packages, e.g. bdvis for biodiversity or rainfreq.

Interactive Data in Practice

There are many R packages that allow for interactive data work in a graphical user interface, including:

- `playwith` - versatile package that works with any graphics function. Graphics can be explored, edited, and exported.
  - requires separate installation of GTK+ on your computer [method varies by OS]
- `rggobi` - powerful 3-D tool for brushing, identifying and manipulating data with a book and online companion site.
  - requires separate installation of GGobi on your computer [method varies by OS]
  - growing long in the tooth [no update since 2012, problematic install on Mac OS X]
- `rattle` - package designed for data mining, includes graphics options alongside other statistical functions
- `latticist` - allows complex linking of plots

Interactive Data on the Web - Rcharts

- Rcharts is a package that uses javascript to create interactive visualizations.
- Lattice-style commands are used.
- The package can output javascript for use in an HTML page.
- Some commands depend on supplemental javascript libraries that must be installed, such as NVD3
- Can embed in documents too, with `slidify`

Interactive Data on the Web - shiny

- The `shiny` package is developed by the Rstudio folks
- You can learn shiny in half a day via the online tutorial
- More custom control of the design is possible with shiny, in comparison to other do-it-all packages
- Graphics use familiar R syntax, with wrappers to implement web functionality
- Every shiny app has the same structure: two R scripts saved together in a directory [ui and server files]
- You must install the shiny server to deliver pages via the web
- For now, we'll just demo a few examples. You can see more in the shiny gallery
- `Rcharts` works with `shiny` too.

Interactive Data on the Web - ggvis

- The `ggvis` package is ALSO developed by the Rstudio folks
- Think `ggplot` meets `shiny`
- Similar syntax to ggplot
- Some ability to add interactive controls

Big Data

- Big data presents special issues for data visualization
- While many techniques and graphics are the same, exploration and plotting must be optimized for the size of the data set
- Representation of the complexity of the data may require special techniques
- `hexbin`
- `bigvis`

Airline Data

For this exercise, we will use Airline on-time performance data.

- This data contains information on every flight in the United States from 1987 to 2008, including arrival and departure times, delays, and other attributes.
- The link above allows access to the full dataset. These are large. For example, the full 2008 data contains 7,009,728 records and is 689 MB.
- We will use an extract of selected variables from January 2008 only. This subset contains 605,765 records and is 37.5 MB.

Binning

- `hexbin` resolves overplotting issues in large data sets by showing density
- There are many other binning methods

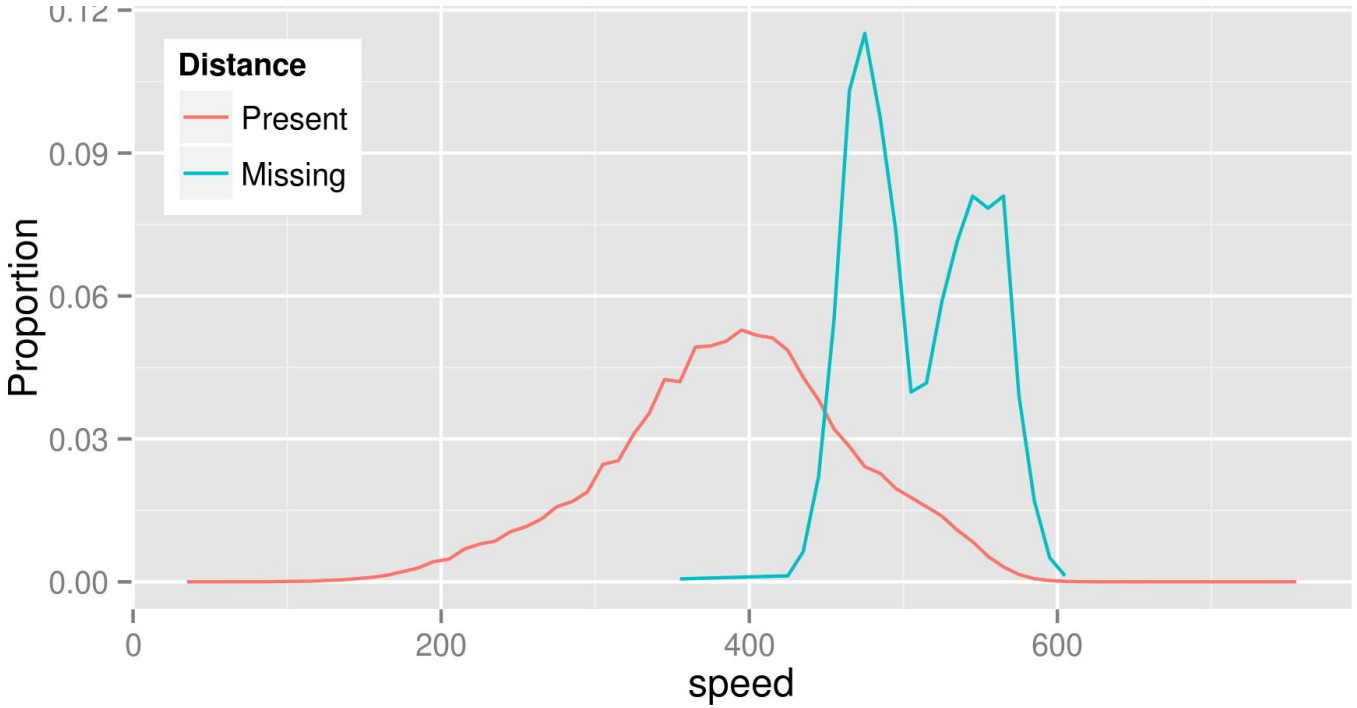
BIGVIS

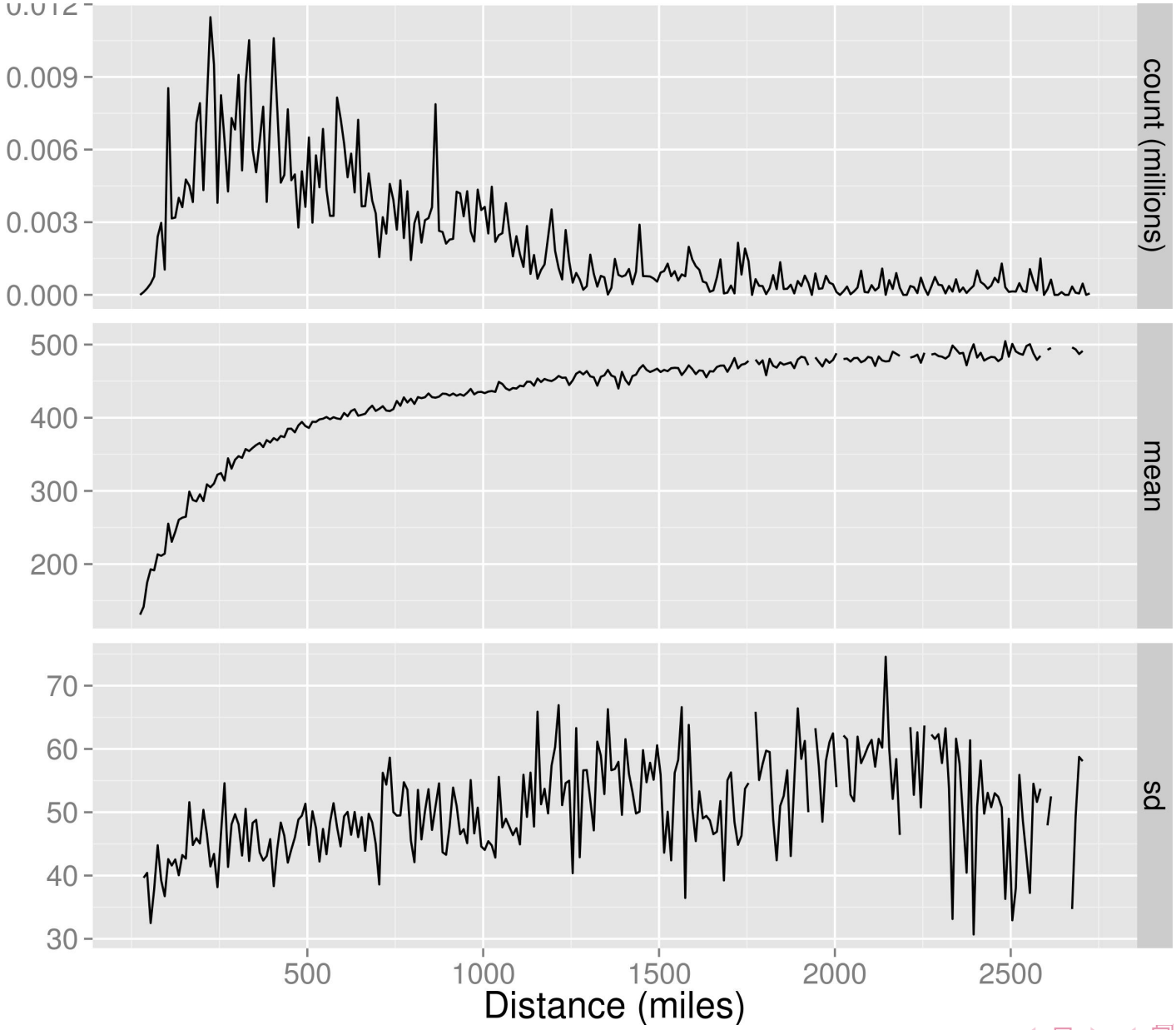
BIGVIS is a relatively new package by Hadley Wickham to deal with the issues of Big Data

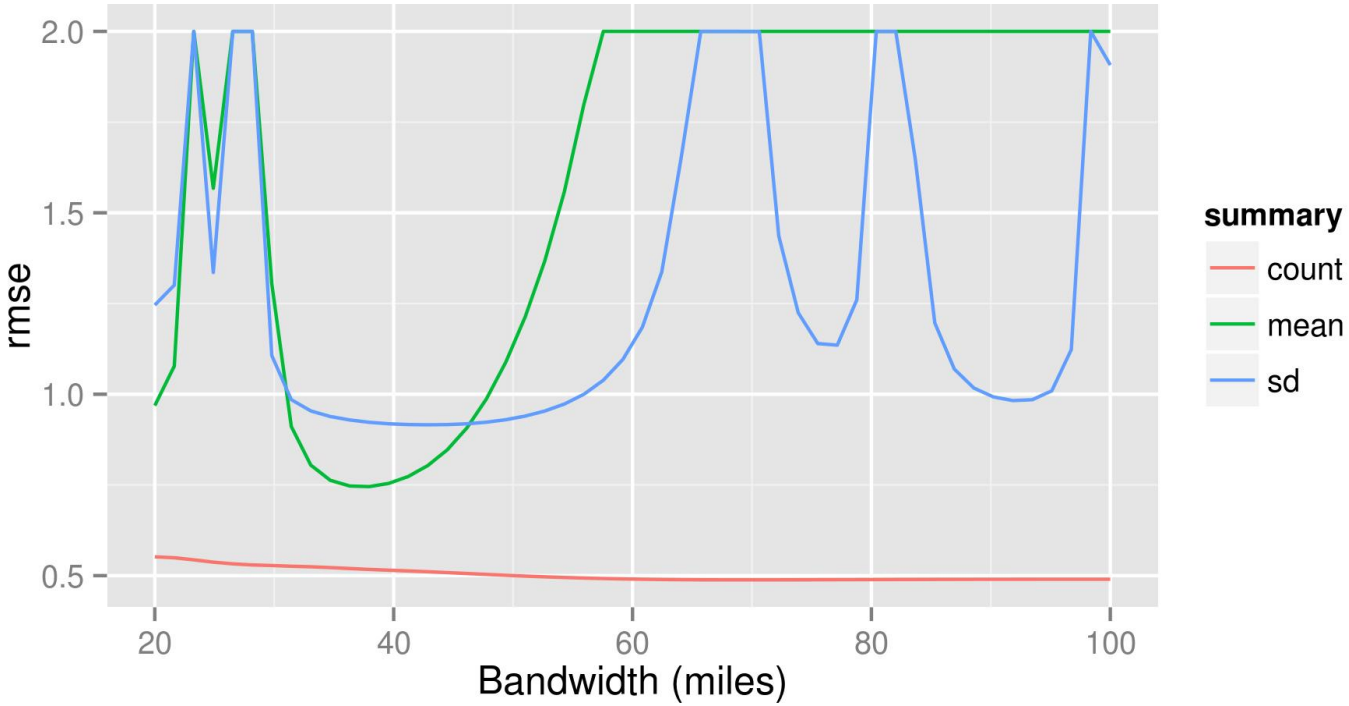
- There is a Preprint and R Meetup presentation by Hadley Wickham
- Complete code, including the extracts adapted for this workshop, is available at https://github.com/hadley/bigvis-infovis
- Target: process 100 million observations in under 5 seconds.
- Fundamental principle: No need for more data points than there are pixels on the screen.

BIGVIS STEPS

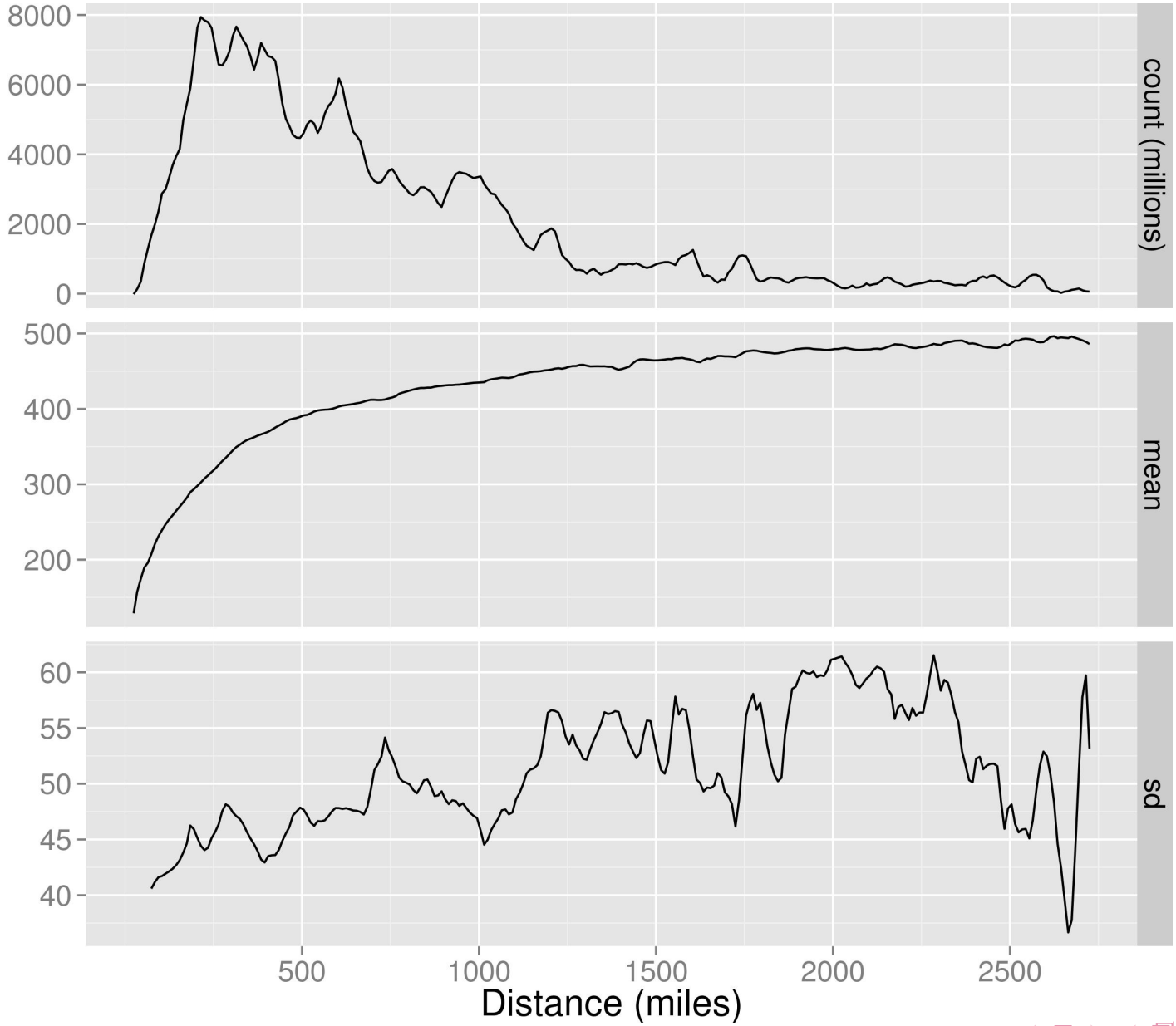
- Condense (`bin`, `condense`)
- Smooth (`smooth`, `best_h`, `peel`)
- Visualize (`autoplot` plus standard methods)
- The following slides are summary output from Hadley's bigvis example.


Distance
Present
Missing
Proportion
0.12
0.09
0.06
0.03
0.00
0
200
400
600
speed

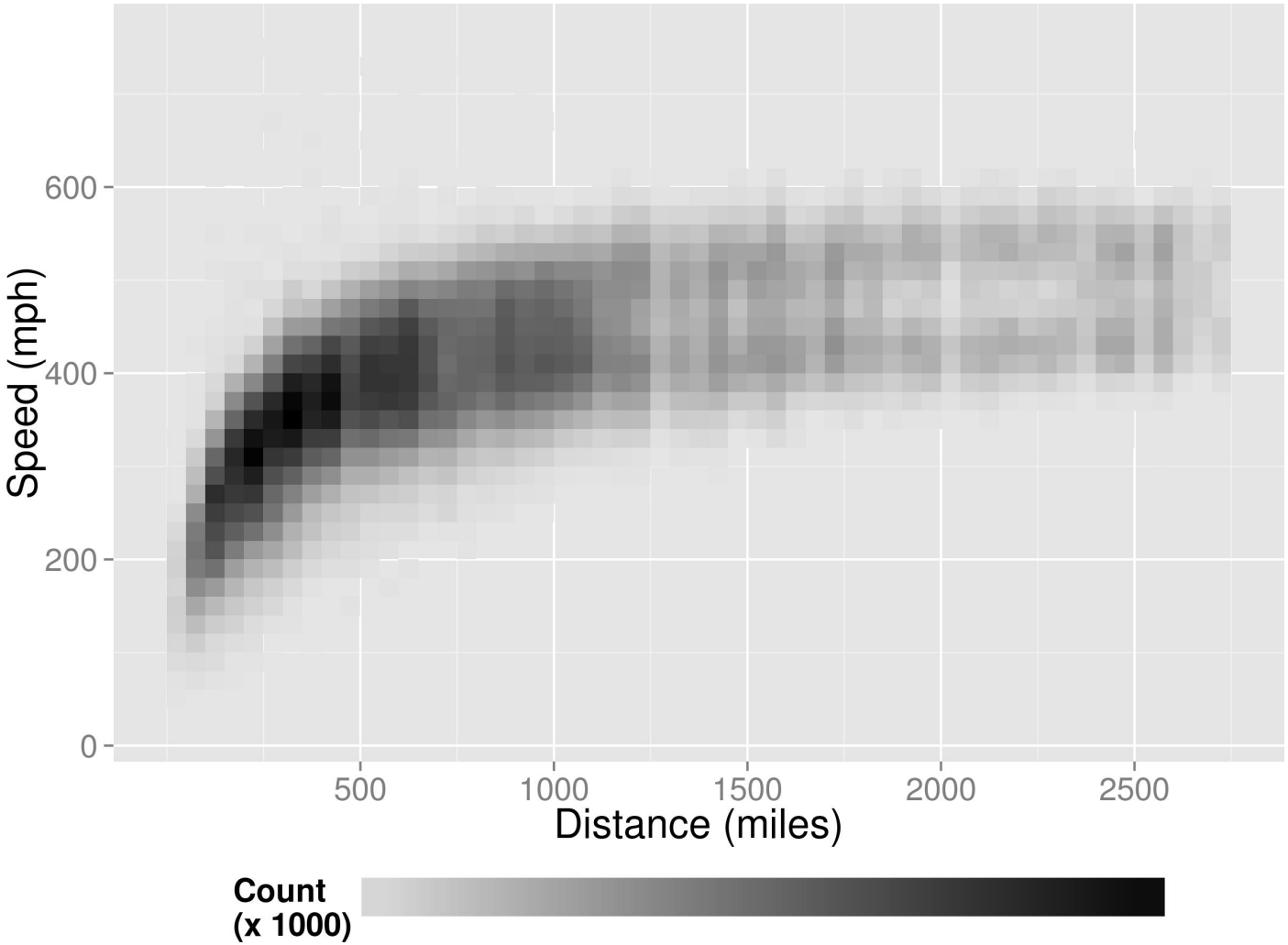

0.012
0.009
0.006
0.003
0.000
count (millions)
500
400
300
200
mean
70
60
50
40
30
sd
500
1000
1500
2000
2500
Distance (miles)


2.0
1.5
1.0
0.5
rmse
20
40
60
80
100
Bandwidth (miles)
summary
count
mean
sd

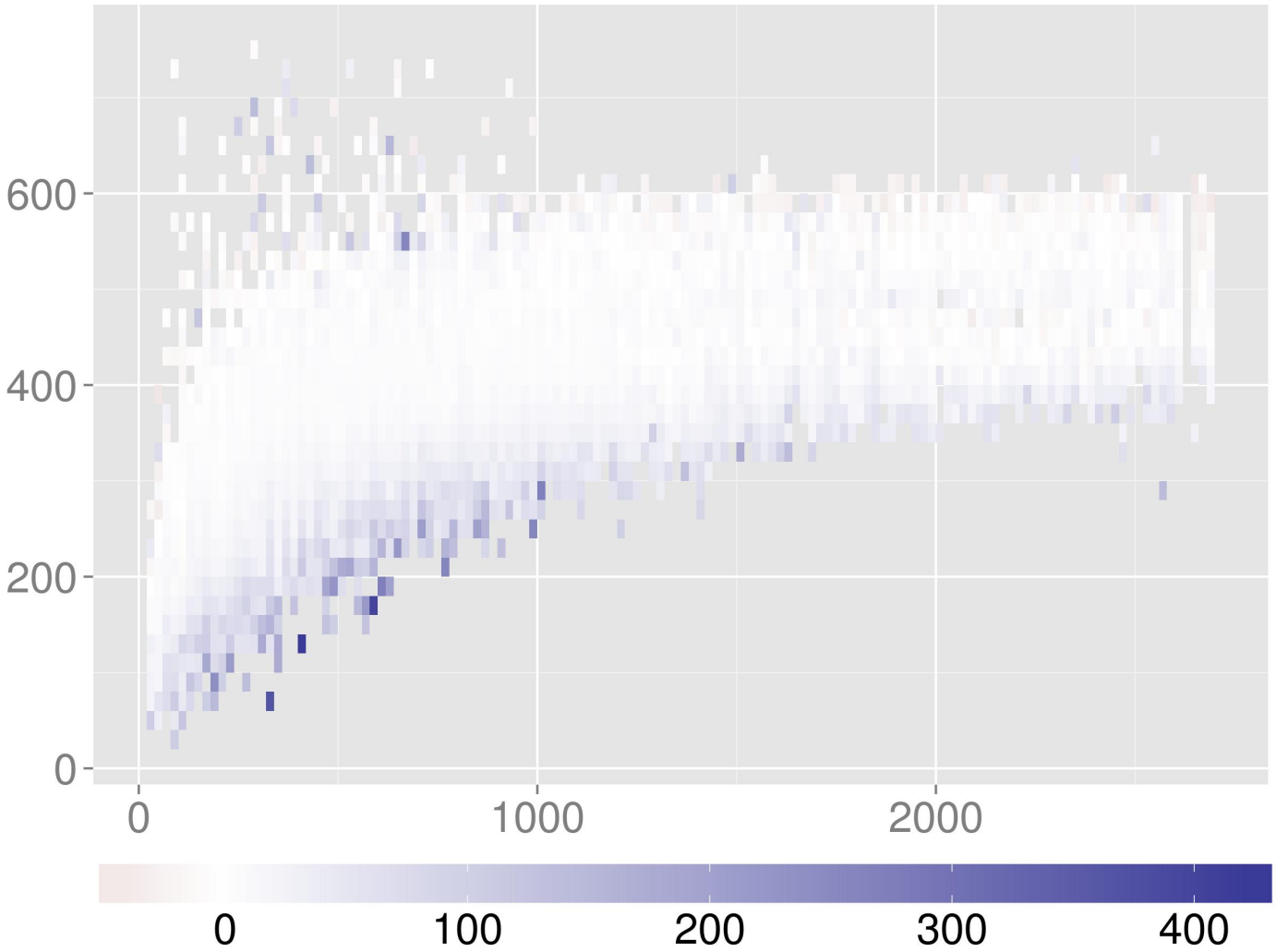
BIGVIS


8000
6000
4000
2000
0
count (millions)
500
400
300
200
mean
60
55
50
45
40
sd
500
1000
1500
2000
2500
Distance (miles)

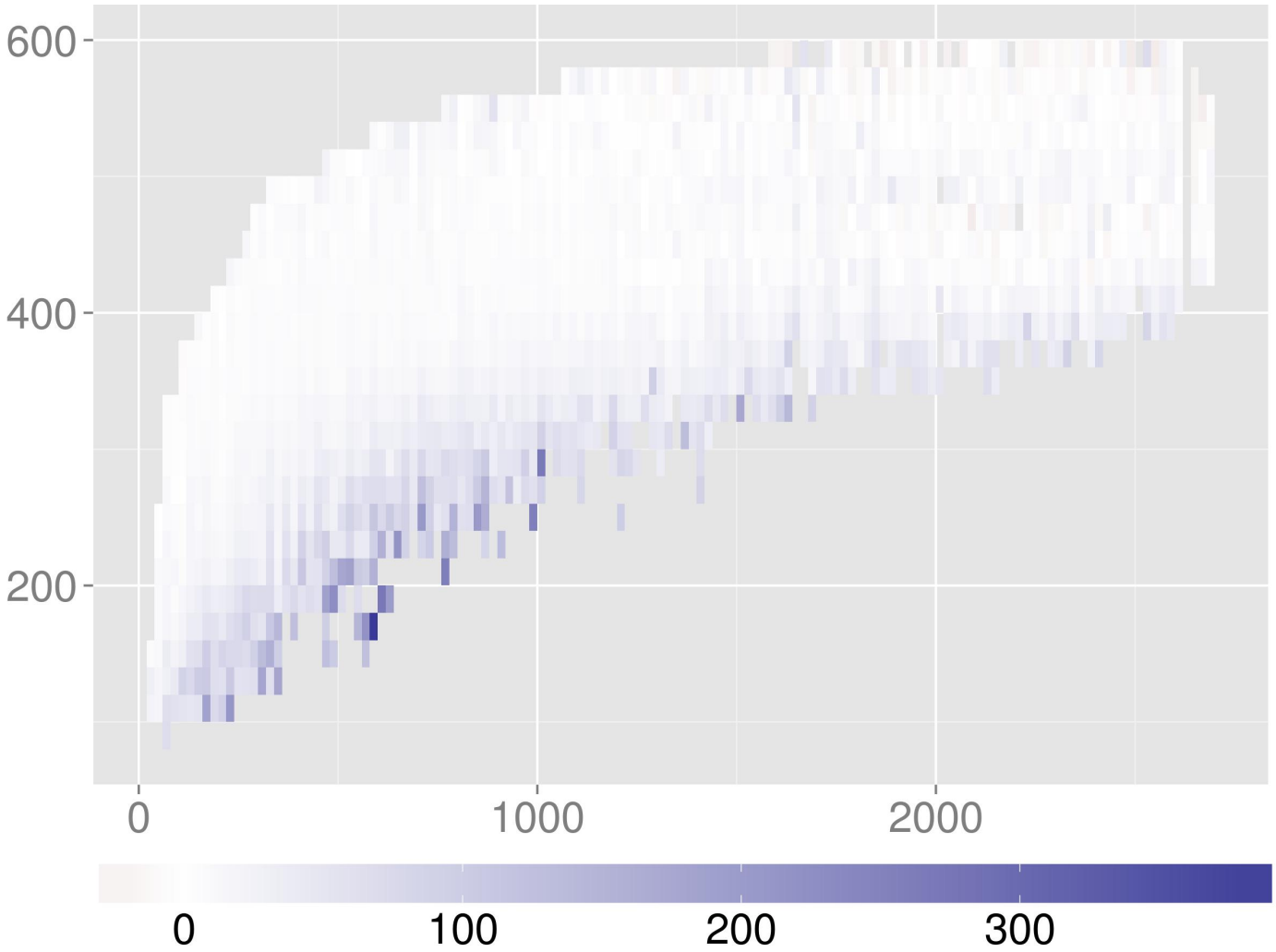
BIGVIS


600
400
200
0
Speed (mph)
500
1000
1500
2000
2500
Distance (miles)
Count
(x 1000)

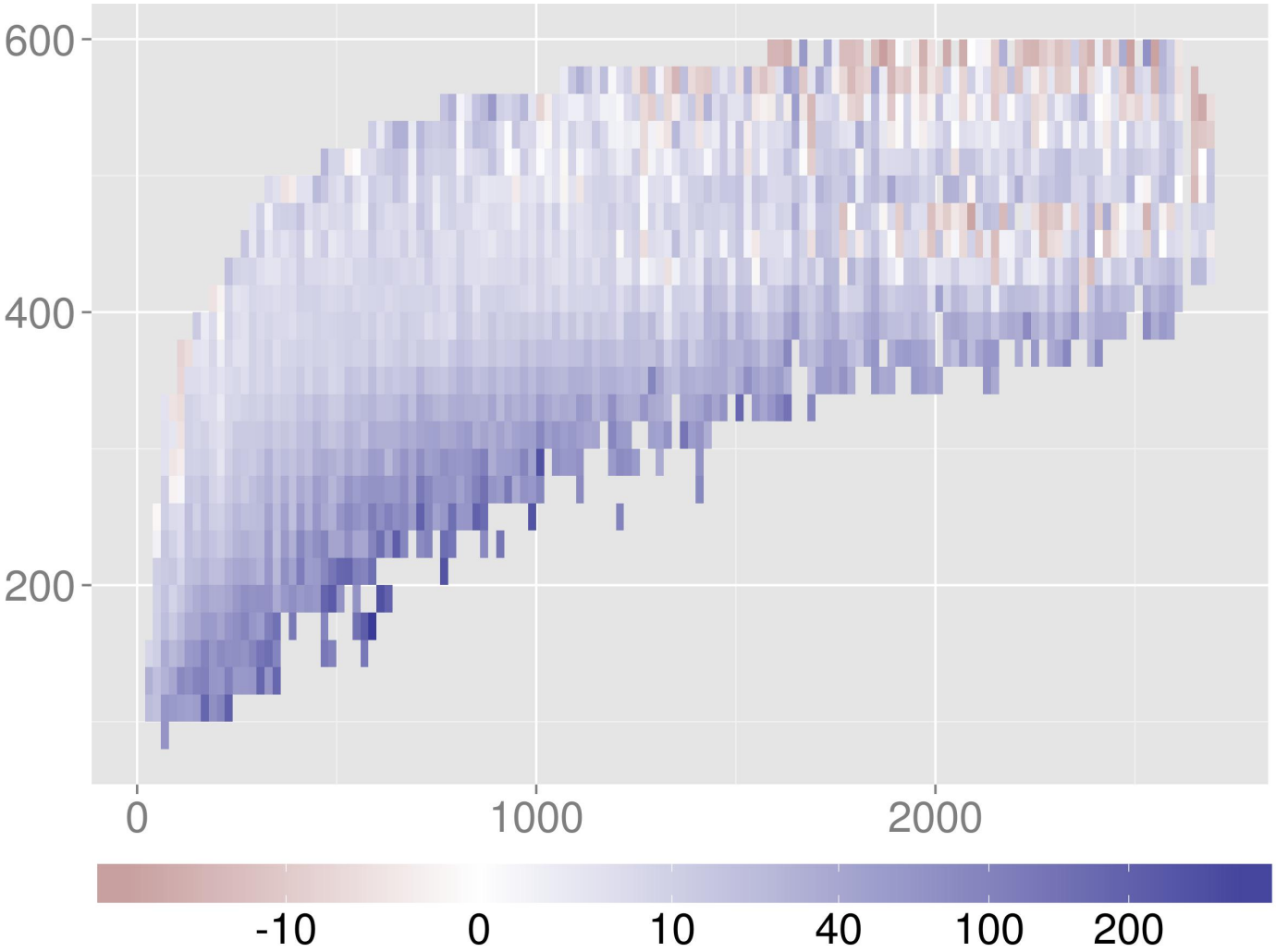
BIGVIS



BIGVIS



BIGVIS



Infographics links

Although not covered here, the following links are a sampling of infographics sites for your later enjoyment:

- Data Storytelling in Video
- Art of Data Visualization - in spite of its title, more on the infographics side
- Parisian Subway Traffic and New York Subway Inequality
- Tulp Interactive
- Information Aesthetics
- Mapping London and London Riots + Twitter
- YouTube Trends Map
- Global Burden of Disease Visualizations

Other considerations

These are not illustrated in the code, but represent future topics for exploration.

- confidence intervals
- missing data [discuss]
- color can be used to indicate certainty
- "scagnostics", borderlining scatterplots (`scagnostics` packages)
- edge blur, crisp vs. fuzzy edges

Keep Exploring

Data Visualization represents a nearly infinite world of possibilty for exploration:

- plunge into programming
- deep dives into data
- indulge in interactivity
- ...have fun and keep learning! [e.g., R-bloggers.com]

References

There is also an online bibliography of references to accompany this presentation on my home page.