Visualization and Data Presentation 2

Jordi Bieger

Reykjavik University
School of Computer Science
Center for Analysis and Design of Intelligent Agents

jbieger@gmail.com

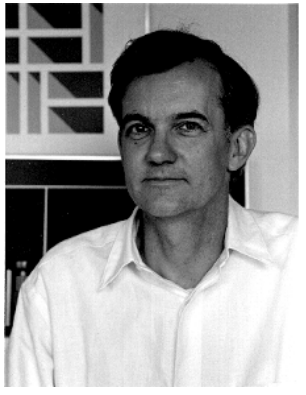
October 20, 2014

Overview

Tufte

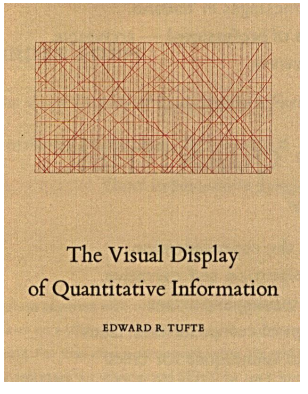
Edward R. Tufte

- Born 1942
- Yale professor of political science, statistics and computer science
- Pioneer of data visualization
- Data visualization pioneer



Theory of data graphics

- The Visual Display of Quantitative Information (VDQI)
- First edition: 1983; Second edition: 1990
- Effectiveness (Graphical Excellence and Integrity)
- Efficiency (Data-Ink, Data Density and Chartjunk)
- Aesthetics



Anscombe

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

$N = 11$
mean of X's $= 9.0$
mean of Y's $= 7.5$
equation of regression line: $Y = 3 + 0.5X$
standard error of estimate of slope $= 0.118$
$t = 4.24$
sum of squares $X - \overline{X} = 110.0$
regression sum of squares $= 27.50$
residual sum of squares of $Y = 13.75$
correlation coefficient $= .82$
$r^2 = .67$

Figure: From Anscombe (1973), "Graphs in Statistical Analysis" via VDQI (page 13)

Anscombe

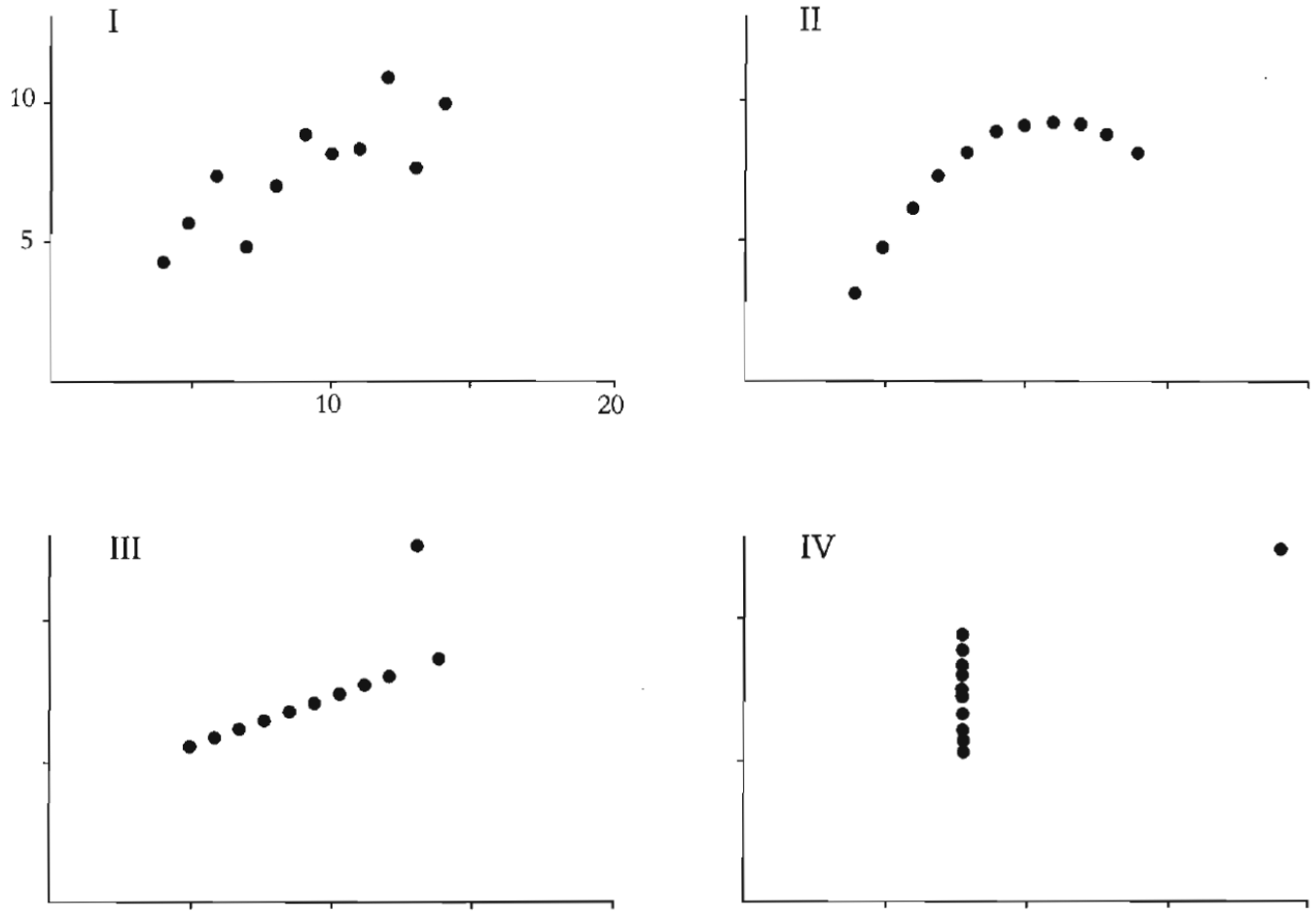
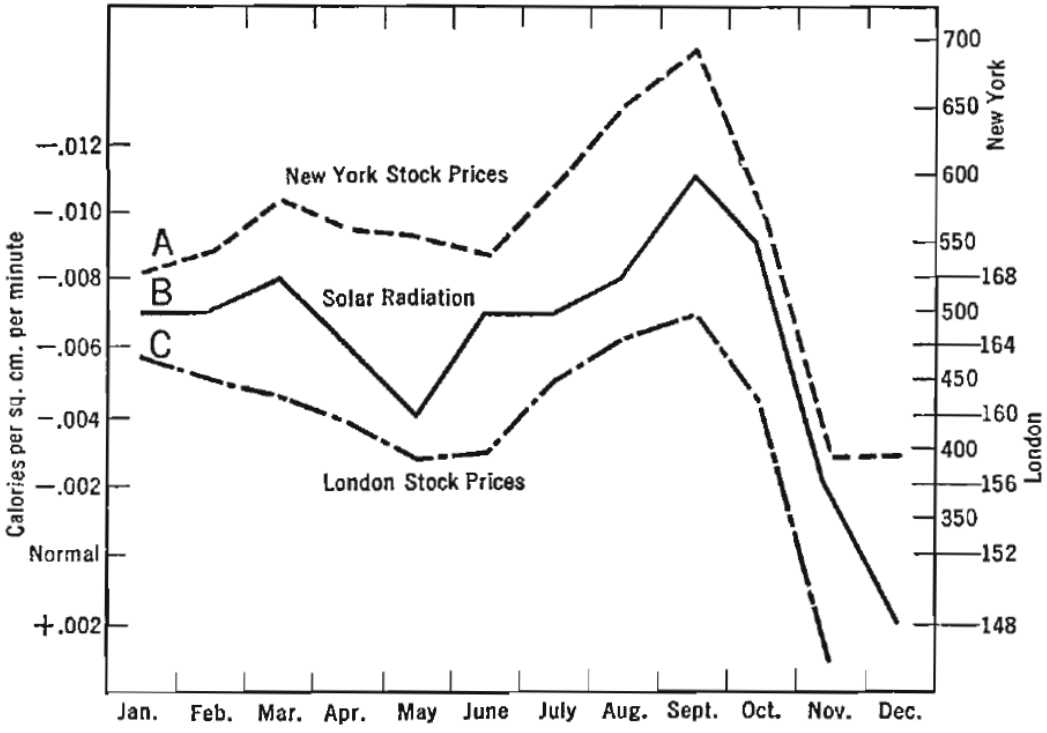


Figure: From Anscombe (1973), "Graphs in Statistical Analysis" via VDQI (page 14)

Multiple Y-Scales



SOLAR RADIATION AND STOCK PRICES

A. New York stock prices (Barron's average). B. Solar Radiation, inverted, and C. London stock prices, all by months, 1929 (after Garcia-Mata and Shaffner).

Figure: From Dewey & Dakin (1947), "Cycles: The science of prediction", p. 144 via VDQI (page 15)

Logarithmic Scales

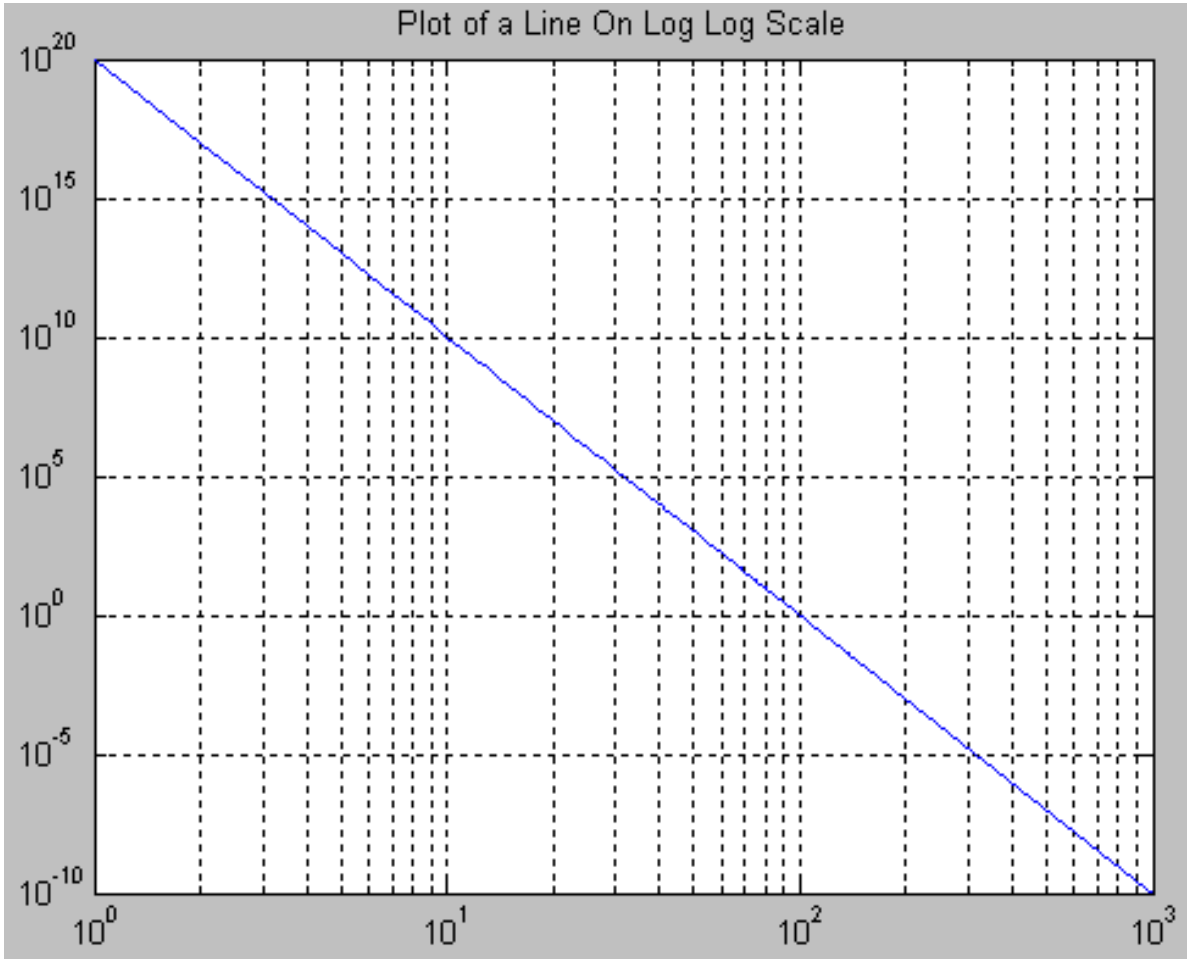


Figure: From **Wikipedia**

Nonlinear Scales

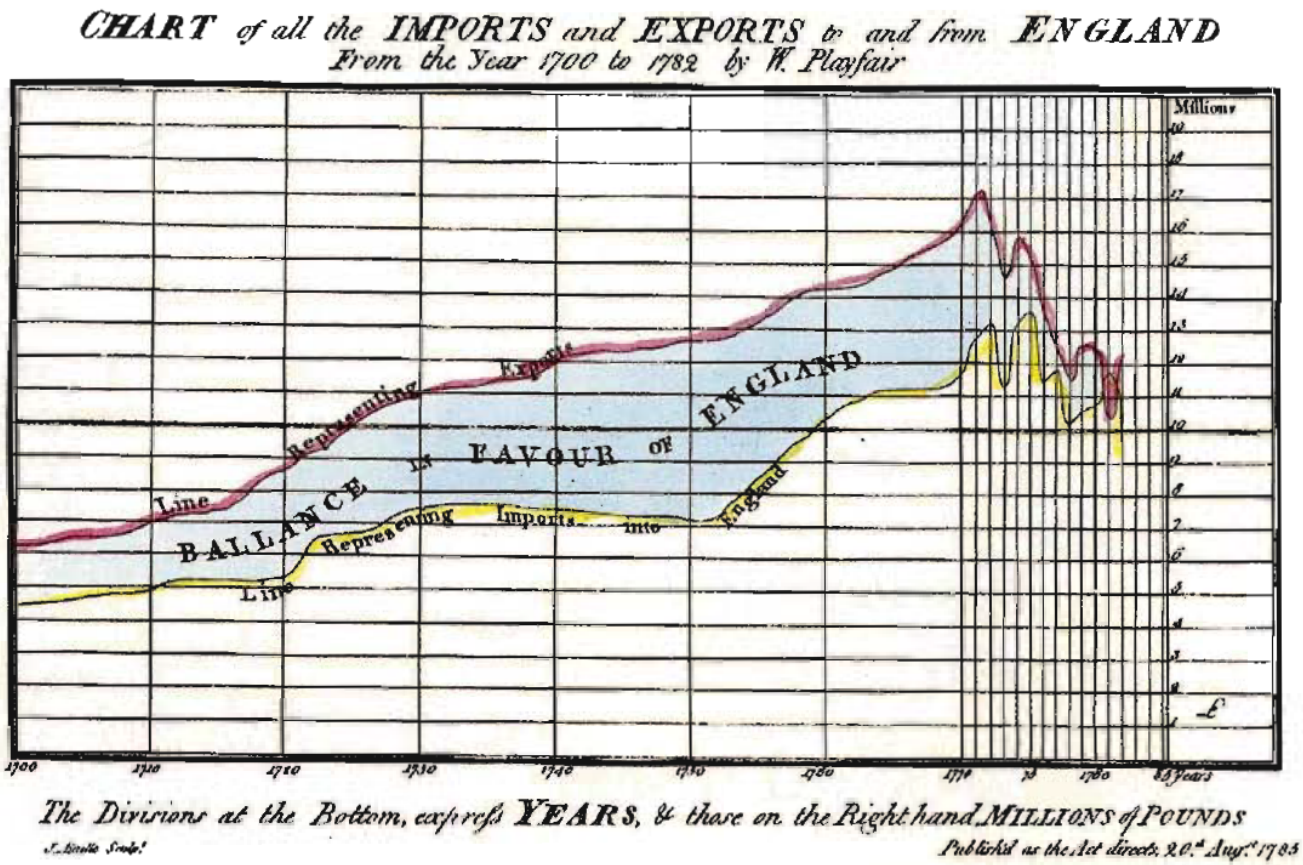


Figure: By Playfair via VDQI (page 32)

Often misleading!

Graphical Integrity

Graphical Integrity

The ability of a graph to provide a visual representation that is consistent with an underlying numerical representation that accurately represents the world.

Subjectivity

Peculiarities of human perception should be taken into account and accommodated rather than exploited. For example, perceived area of a circle $=$ (actual area)^x where $x = .8 \pm .3$.

Lie Factor

$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

Exercise

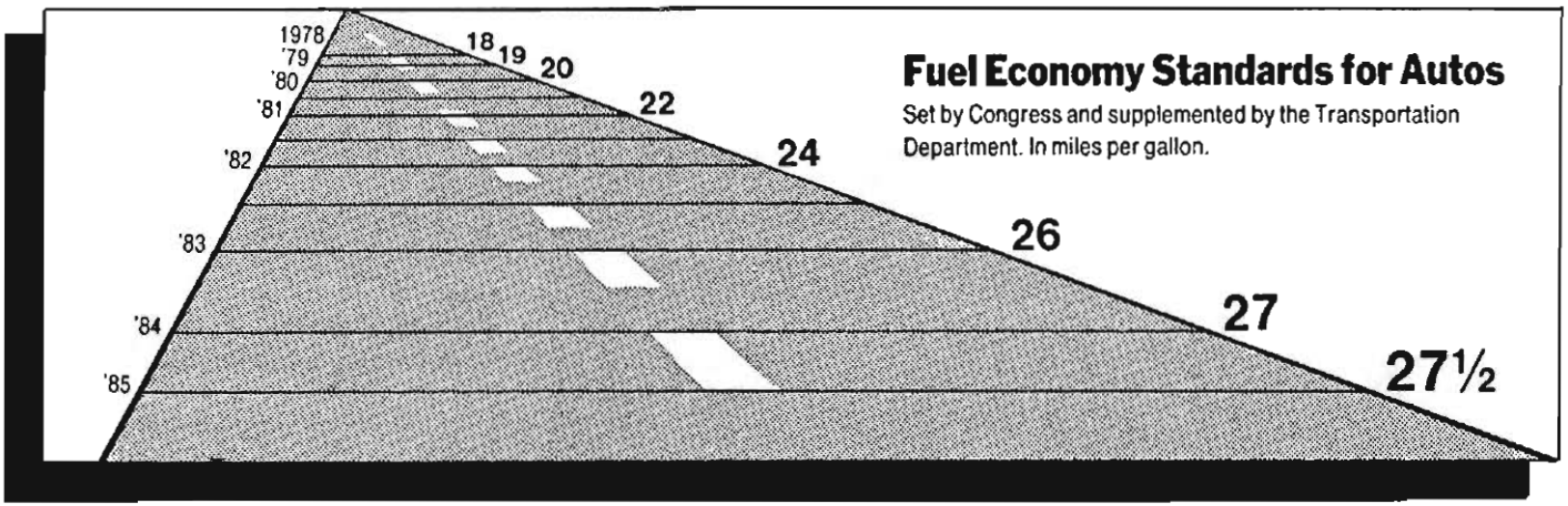


Figure: Adapted from New York Times, August 9 1978, p. D-2 via VDQI (page 57)

Exercise

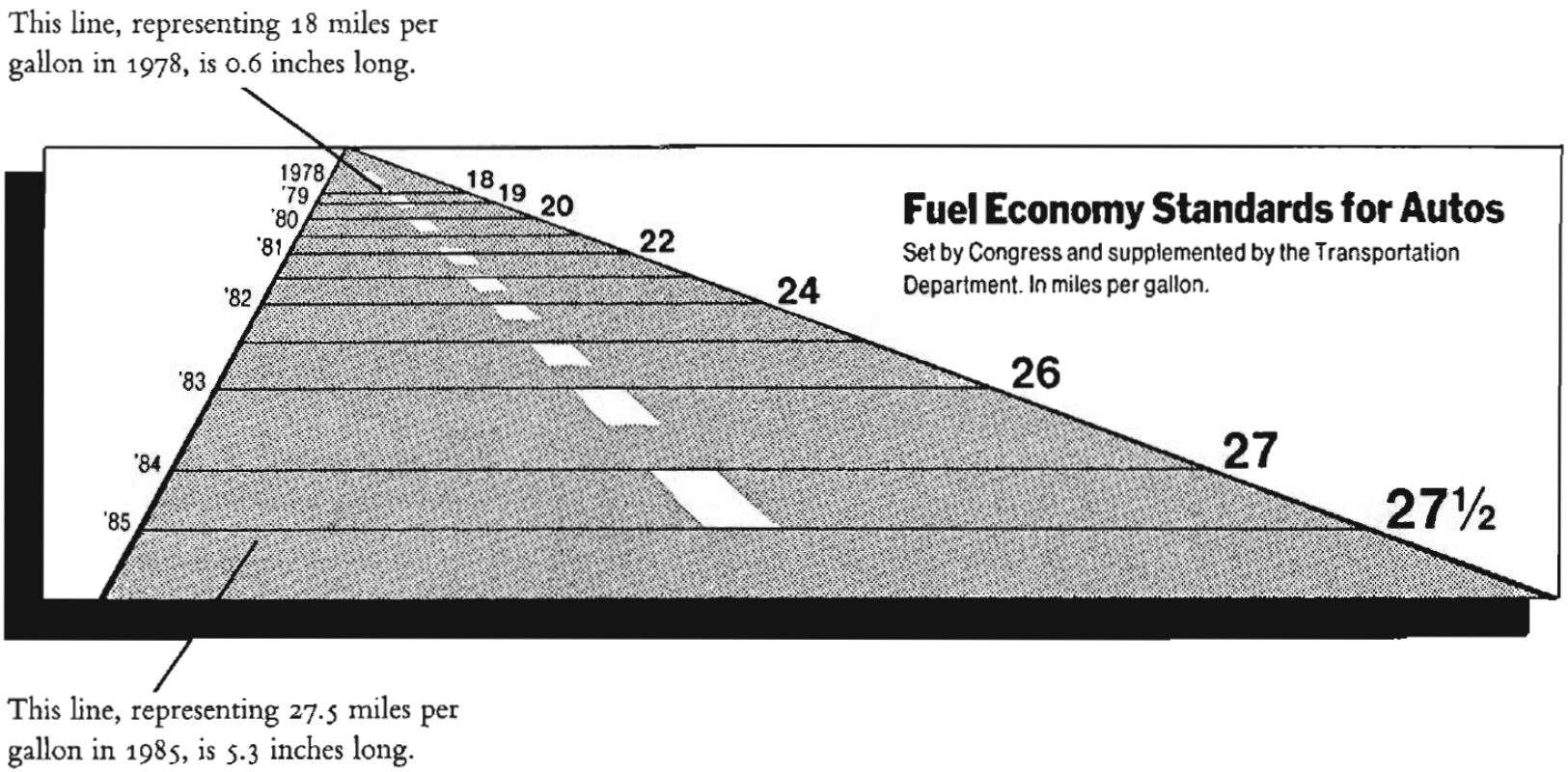


Figure: From New York Times, August 9 1978, p. D-2 via VDQI (page 57)

Design and Data Variation

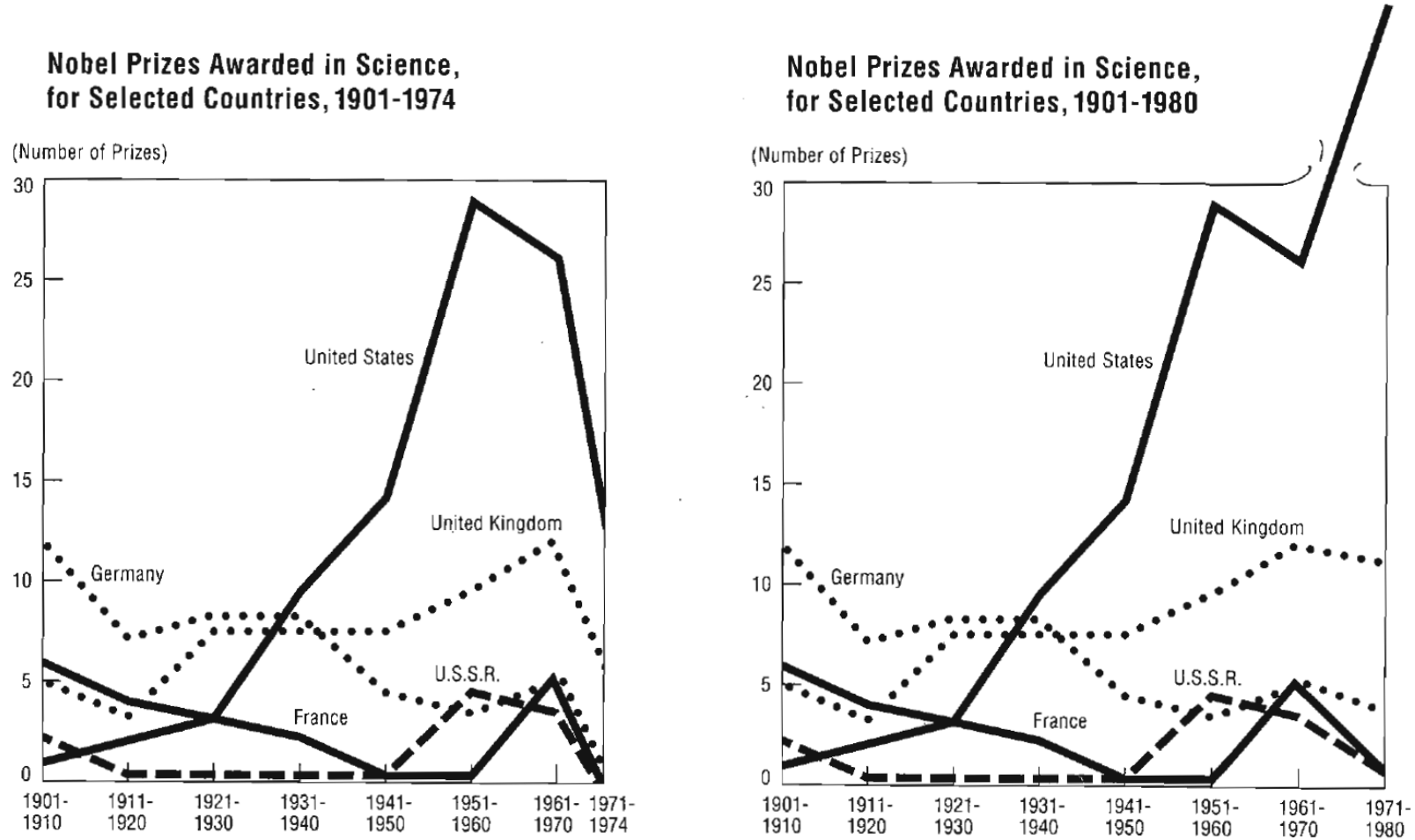


Figure: From National Science Foundation, Science Indicators, 1976 (Washington D.C., 1976) via VDQI (page 60)

2-D representation of 1-D data

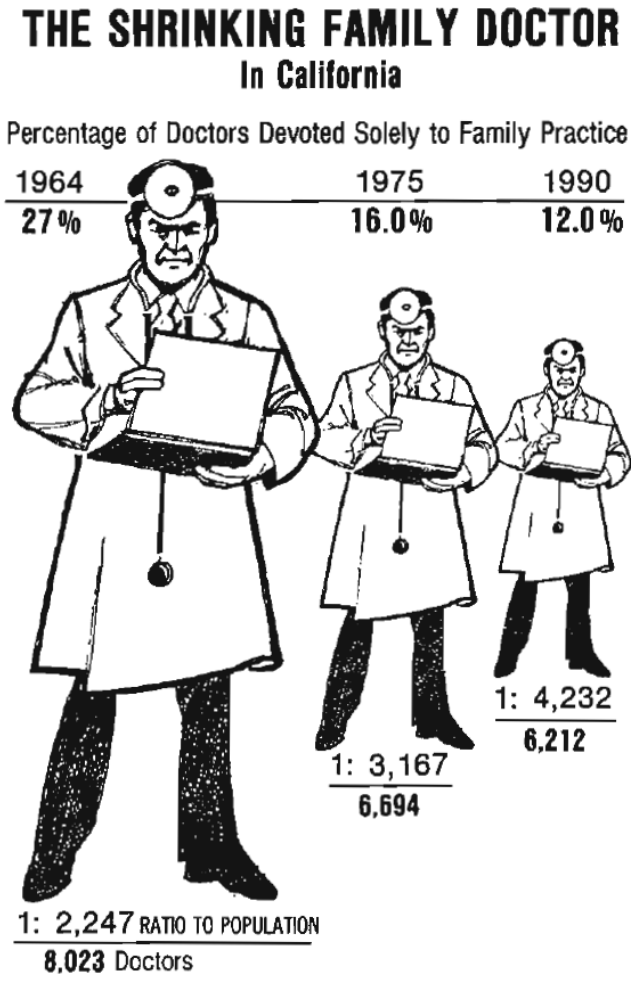


Figure: From Los Angeles Times, August 5 1979, p. 3 via VDQI (page 69)

3-D representation of 1-D data

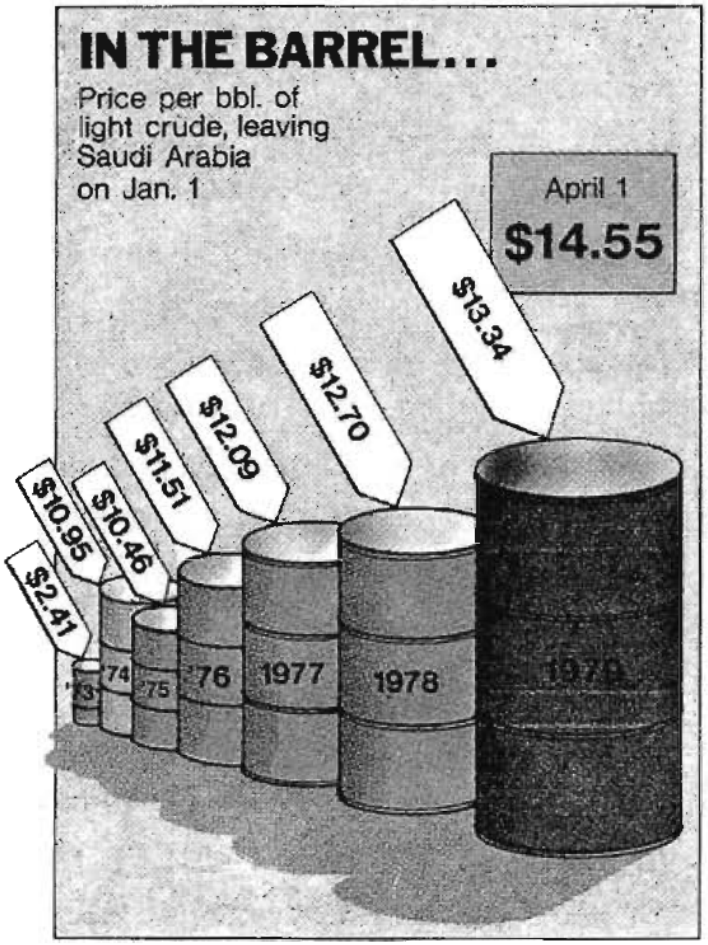


Figure: From Time, April 9 1979, p. 57 via VDQI (page 62)

Money

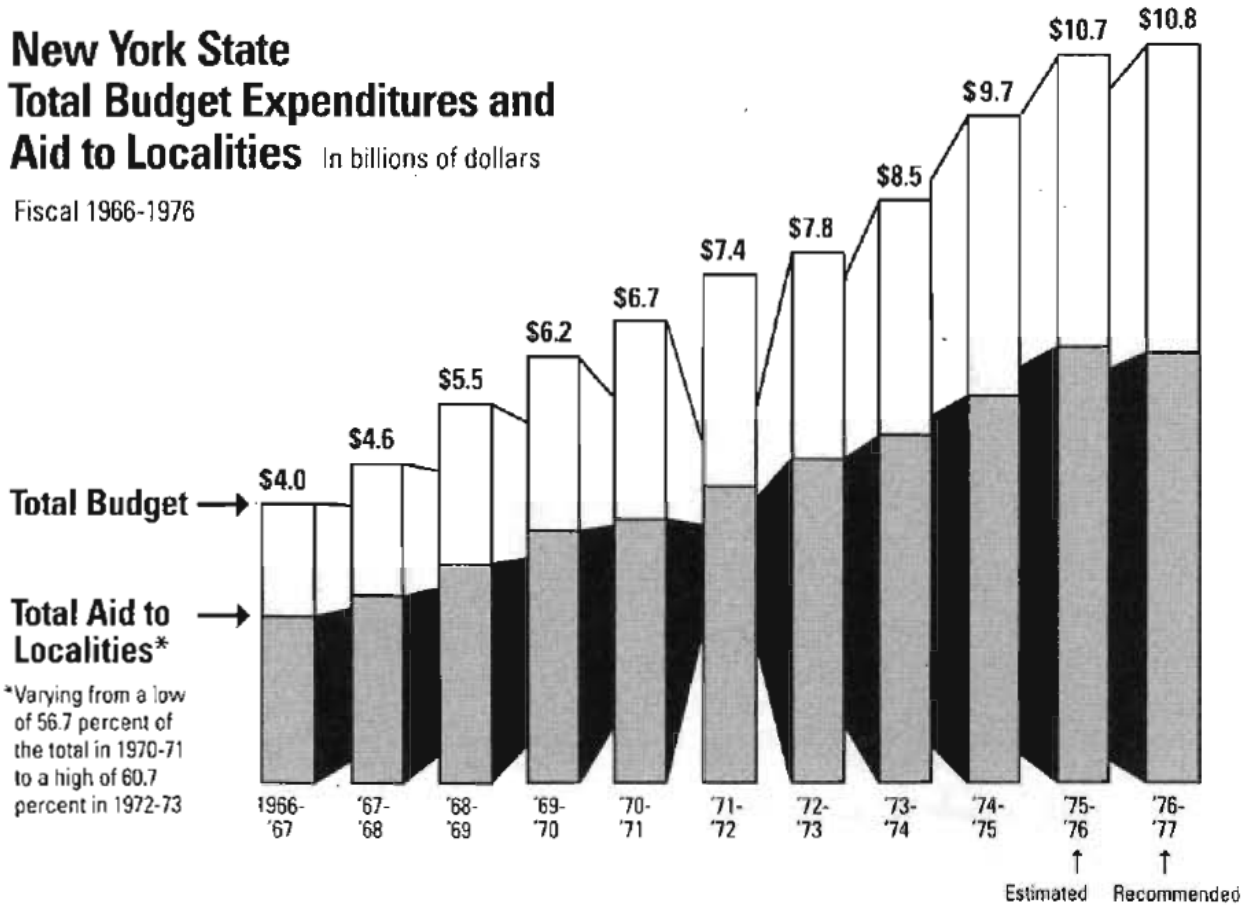


Figure: From New York Times, February 1 1976, p. IV-6 via VDQI (page 66)

Correct for inflation and other factors

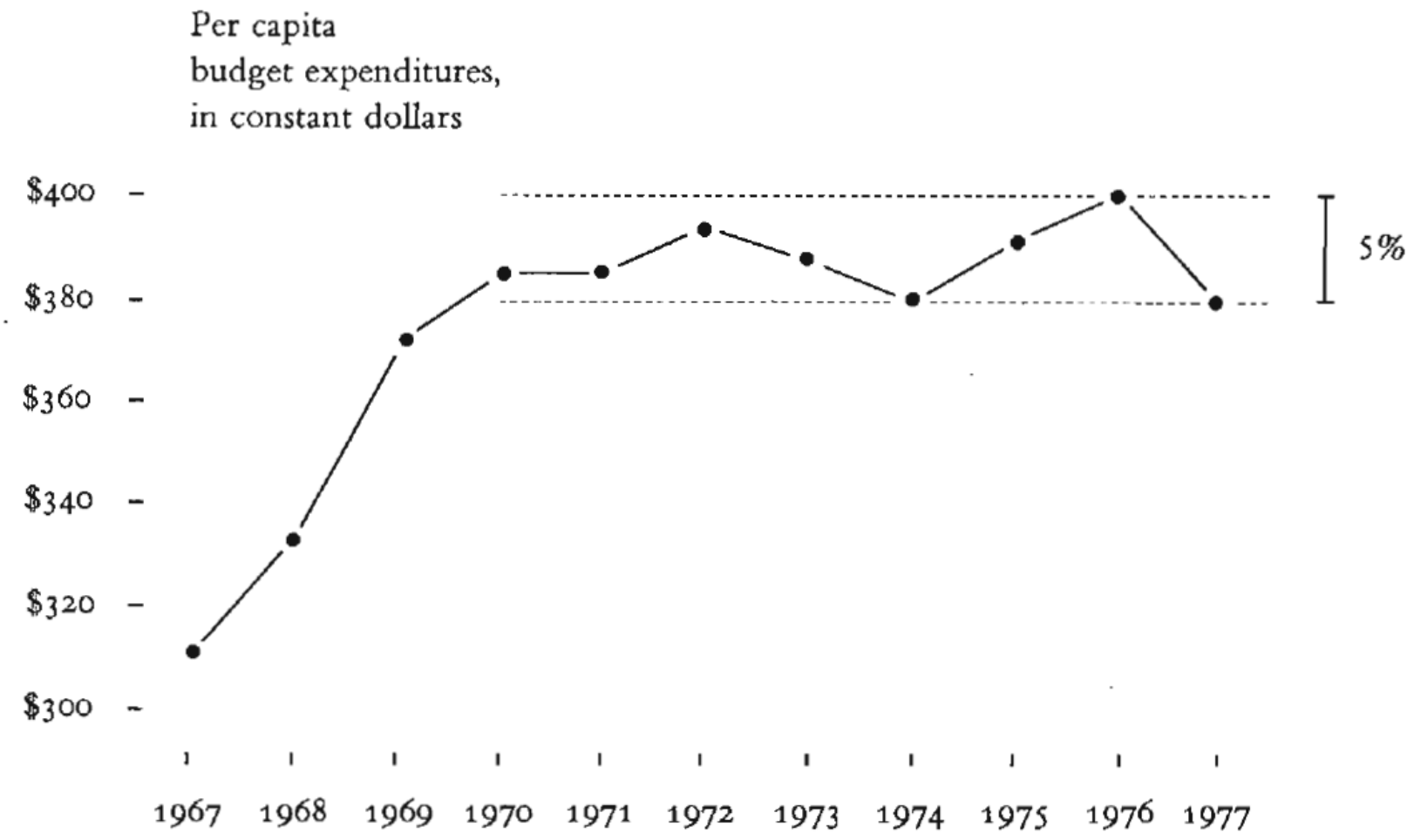


Figure: From VDQI (page 68)

Context

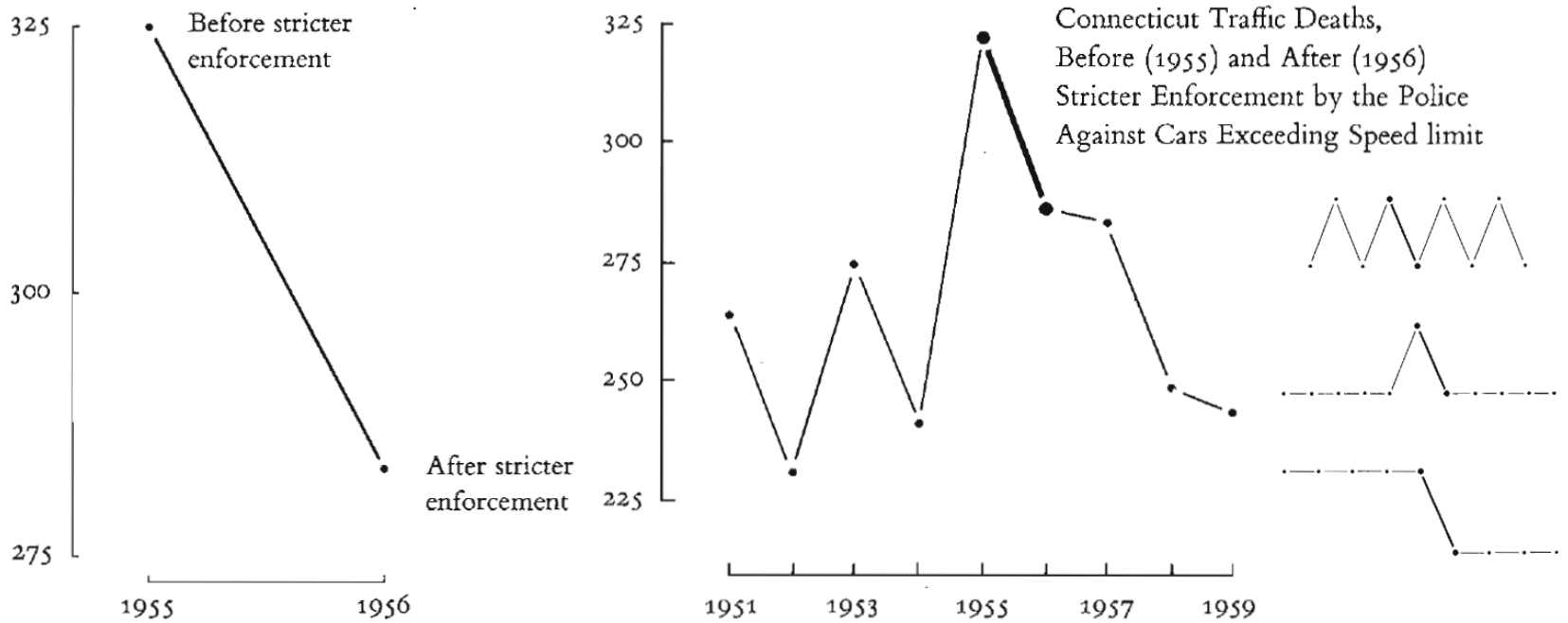


Figure: From Campbell & Ross (1970), "The Connecticut Crackdown on Speeding: Time Series Data in Quasi-Experimental Analysis" via VDQI (page 74)

Context

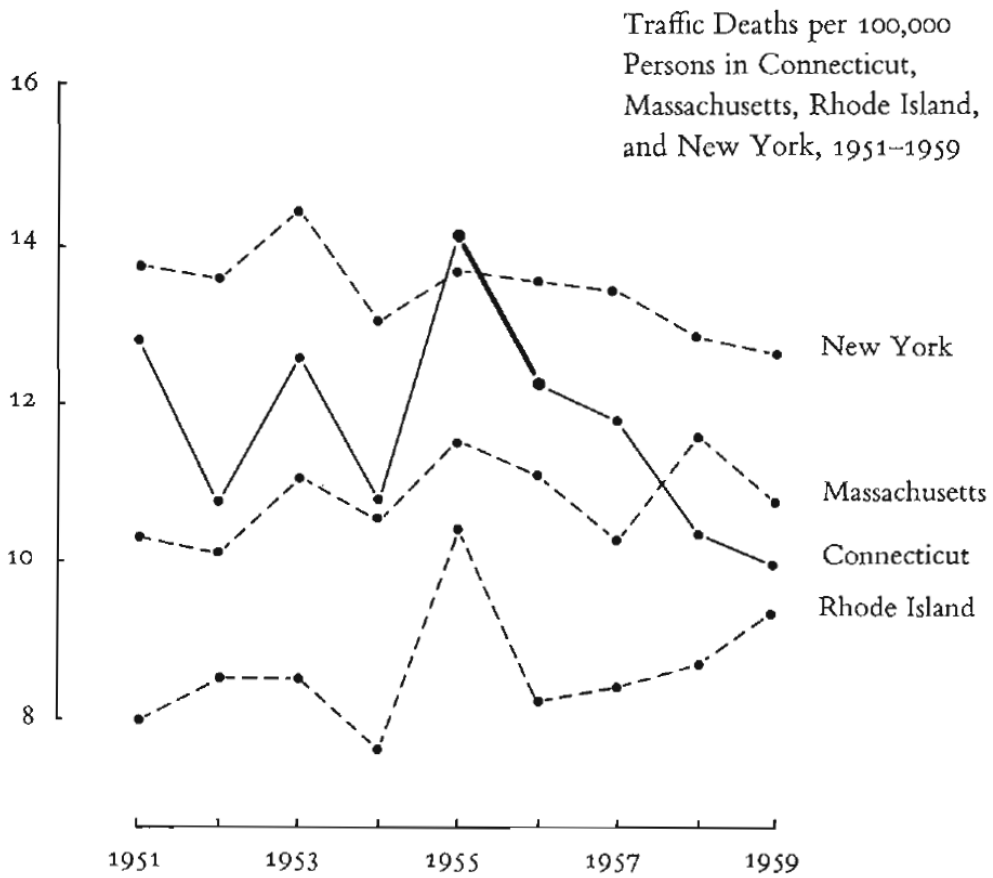
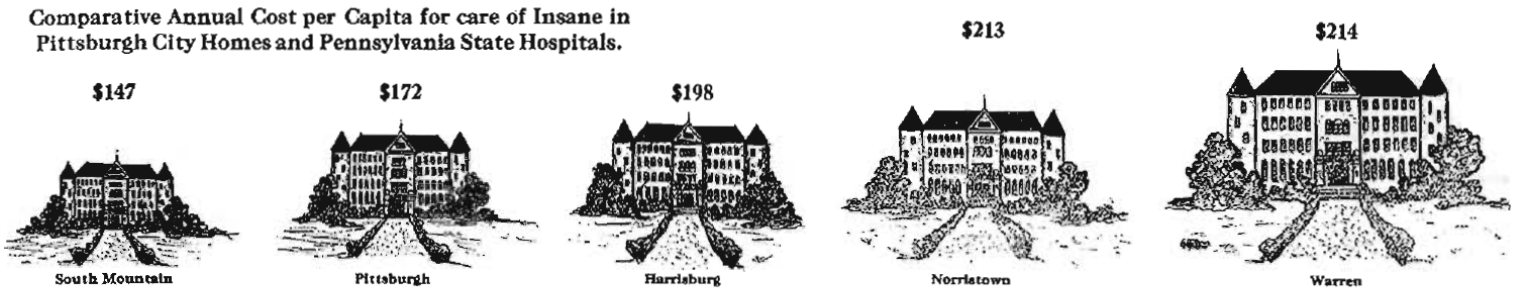


Figure: From Campbell & Ross (1970), "The Connecticut Crackdown on Speeding: Time Series Data in Quasi-Experimental Analysis" via VDQI (page 75)

Numbers have order and magnitude



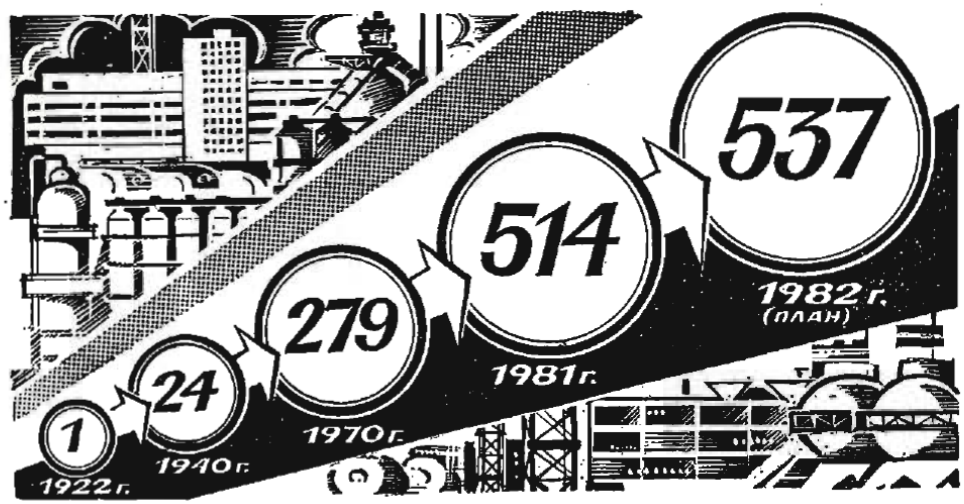


Figure: From Pittsburgh Civic Commission (1911), "Report on Expenditures of the Department of Charities" and Pravda, May 24 1982 p.2 via VDQI (page 55 and 76)

Data-Ink

Data-ink

Data-ink is the non-erasable core of a graphic; the non-redundant ink arranged in response to variation in the numbers presented.

Data-Ink Ratio

$$\begin{aligned}\text{Data-ink ratio} &= \frac{\text{data-ink}}{\text{total ink used to print the graphic}} \\ &= \text{proportion of a graphic's ink devoted} \\ &\quad \text{to the non-redundant display of data-information} \\ &= 1.0 - \text{proportion of a graphic that can be erased} \\ &\quad \text{without loss of data-information.}\end{aligned}$$

Examples

Data-Ink

- Lines in a line graph, bars in a bar graph, dots in a scatter plot, etc.
- Labels
- Data values

Non-Data-Ink

- Axes
- Ticks
- Grid lines
- Decorations

Maximize Data-Ink-Ratio

- Depict more data
- Erase non-data-ink
- Erase redundant data-ink

Within reason!

Exercise

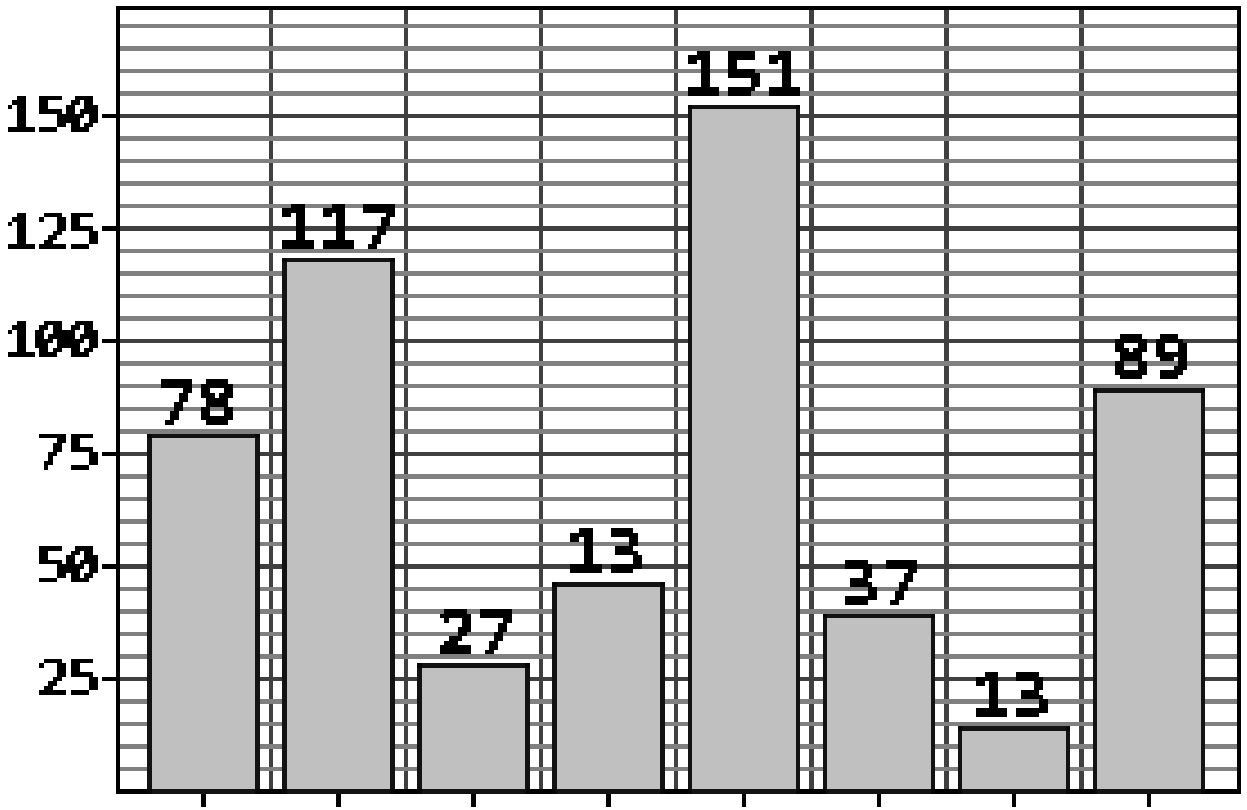


Figure: See VDQI (page 96 and 126-128)

How can we increase the data-ink-ratio?

Exercise 2

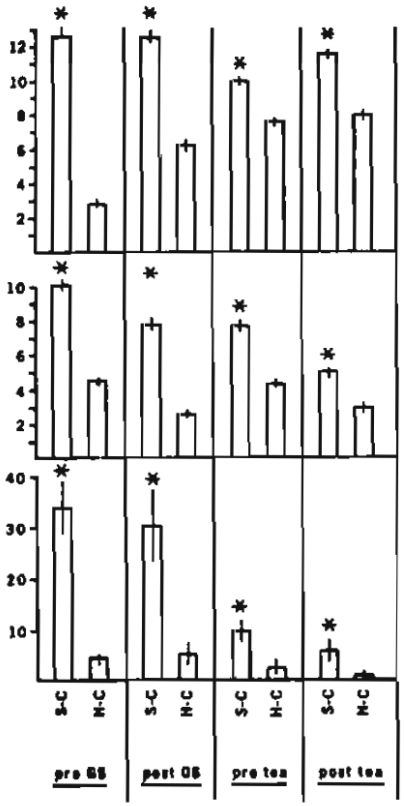


Figure: From Kuznicki & McCutcheon (1979) via VDQI (page 100)

Exercise 2

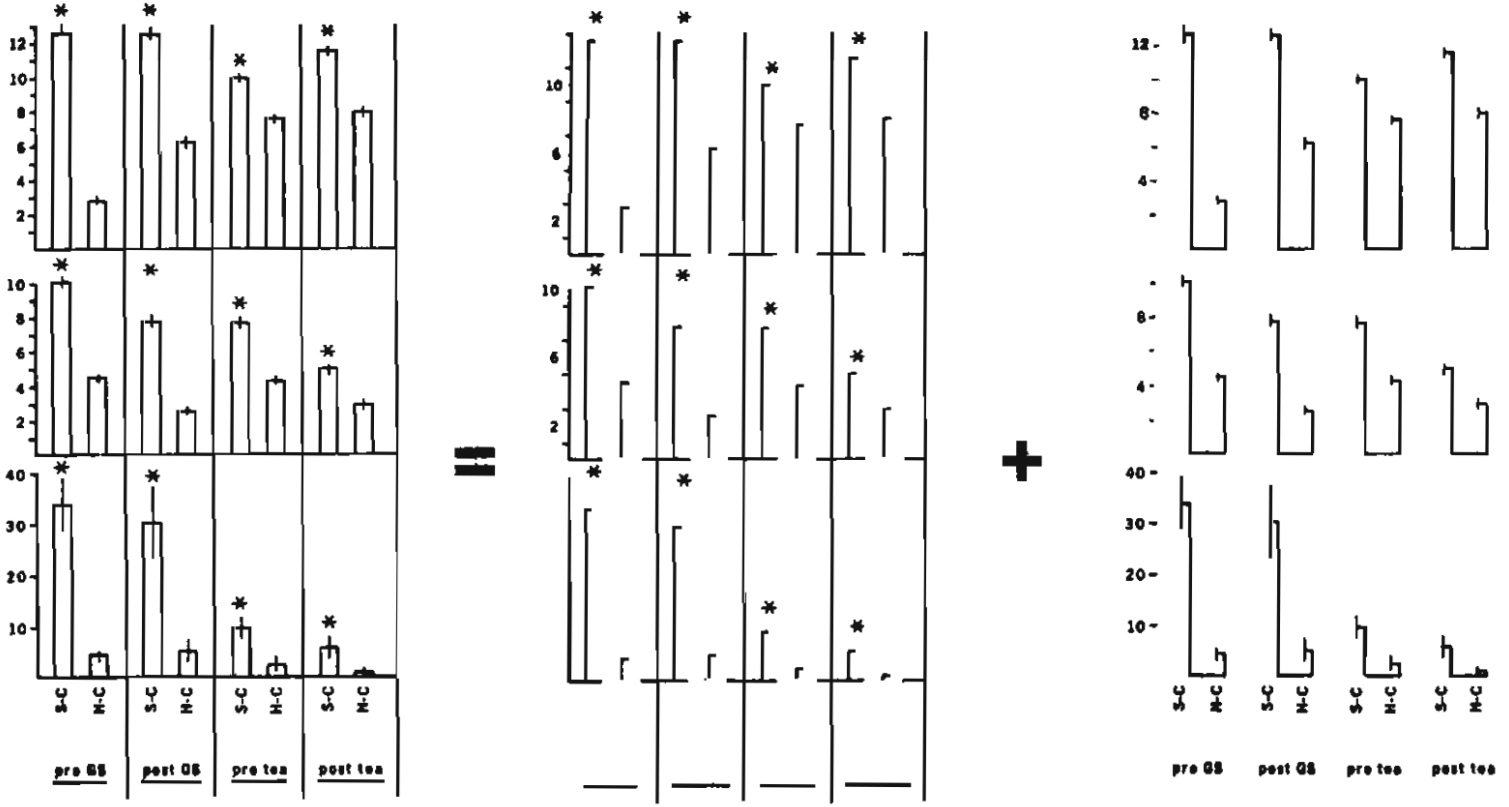


Figure: From VDQI (page 102)

Sparklines

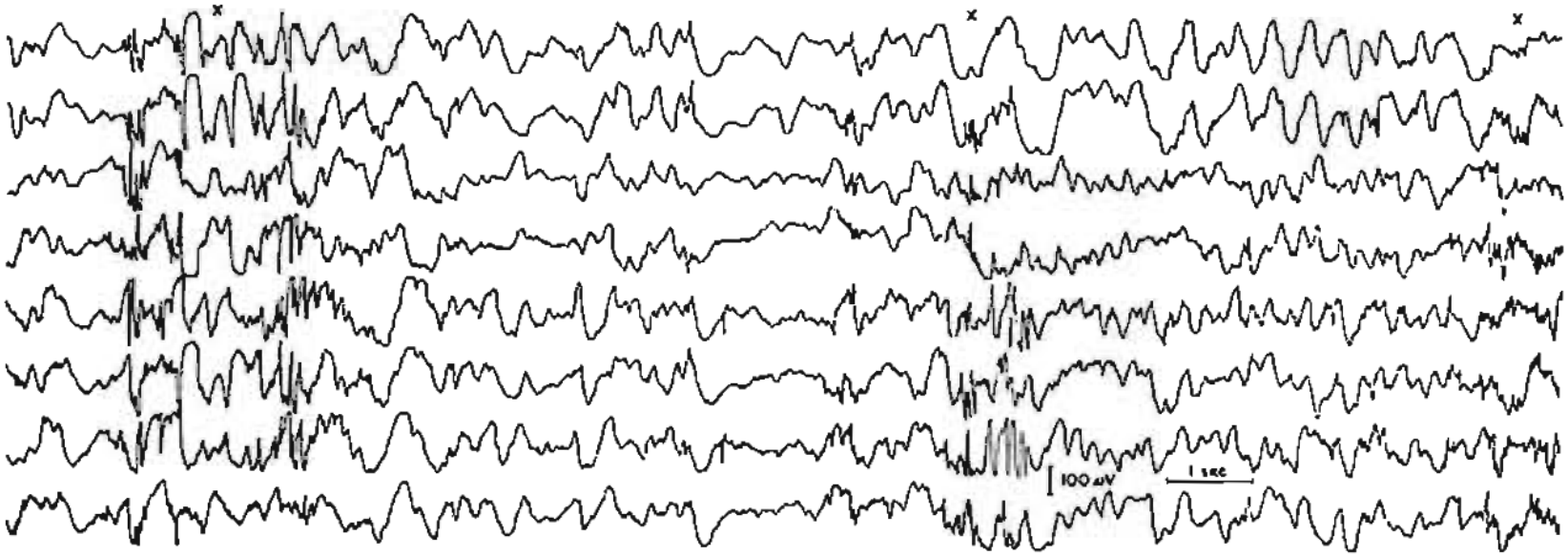


Figure: From Kooi (1971), "Fundamentals of Electroencephalography" via VDQI (page 93)

Boxplots

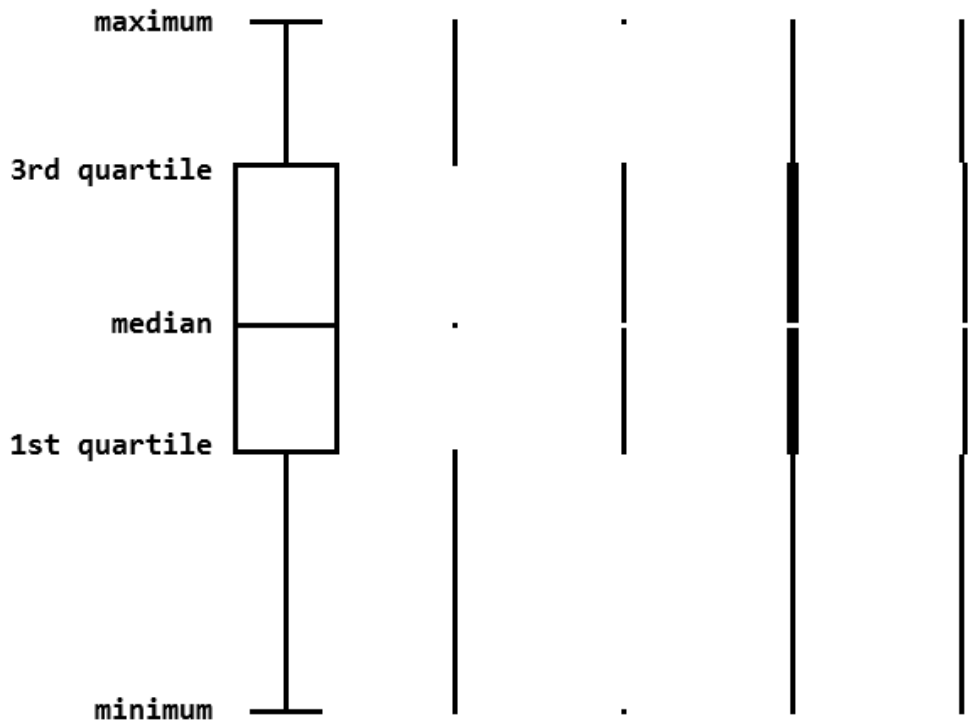


Figure: After VDQI (page 123-125)

Range-Frame

Figure: From VDQI (page 132)

Range-Frame

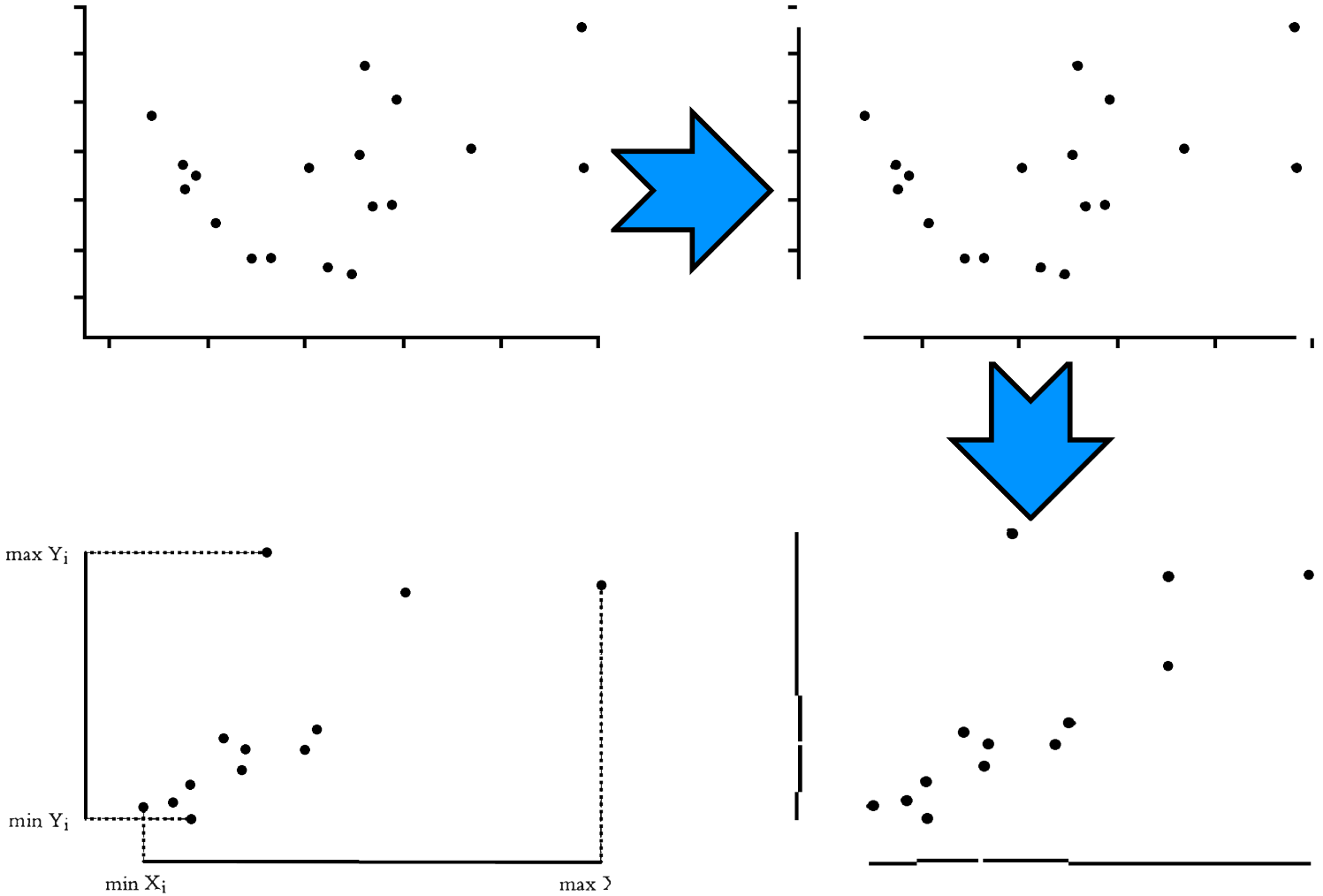


Figure: From VDQI (page 130-132)

Dot-Dash-Plot

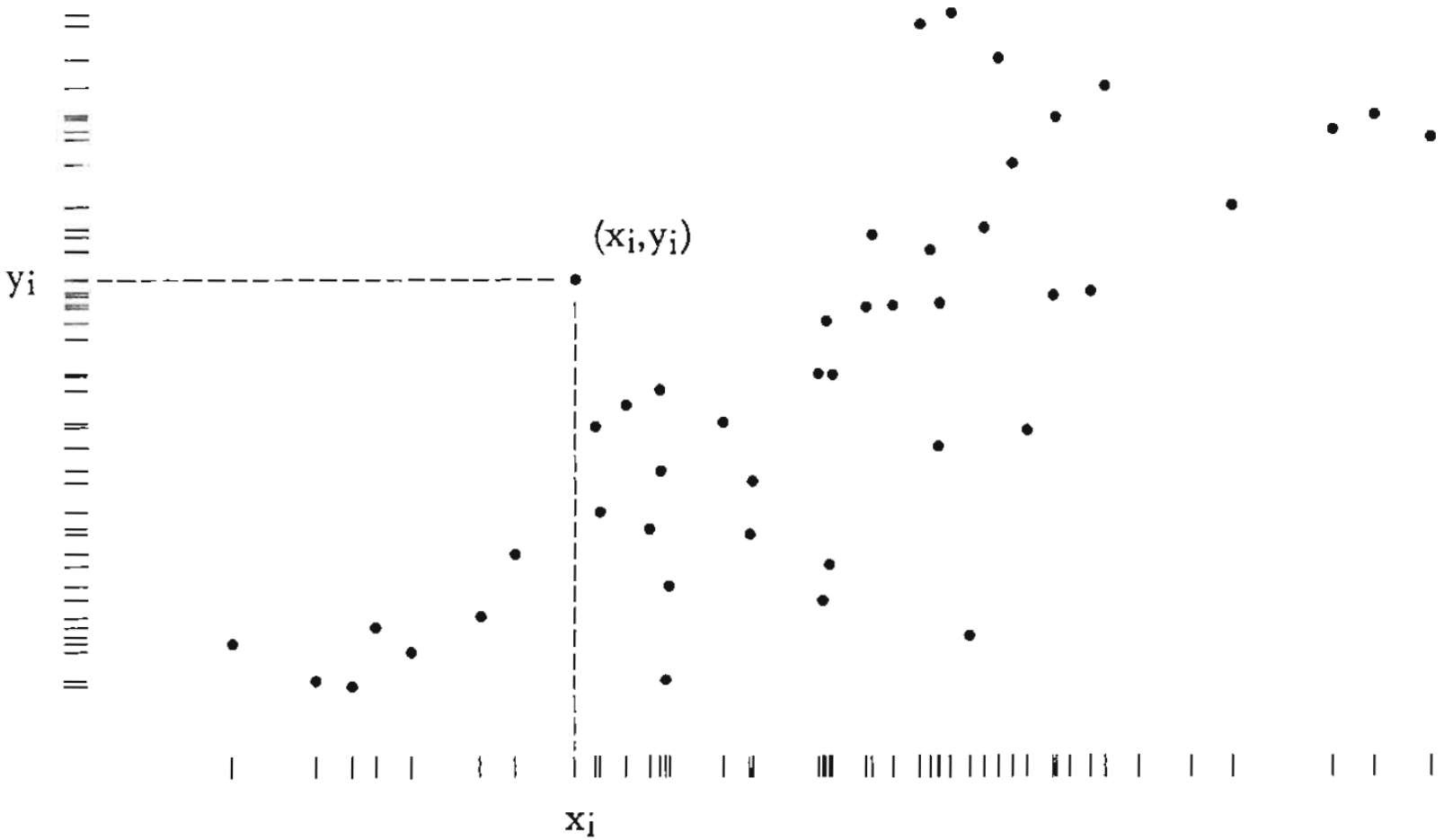


Figure: From VDQI (page 133)

Rugplot

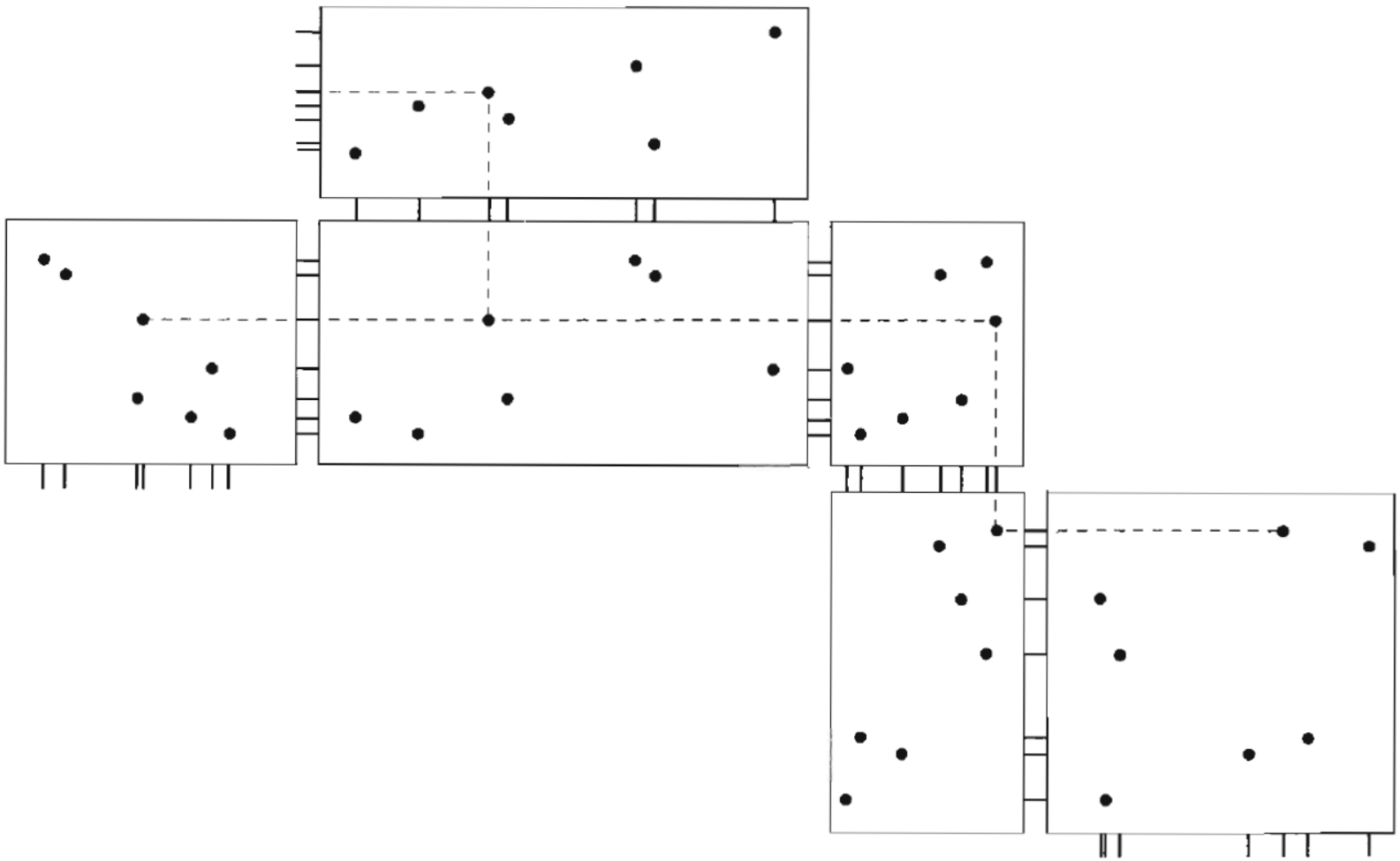


Figure: From VDQI (page 135)

Distribution on axes

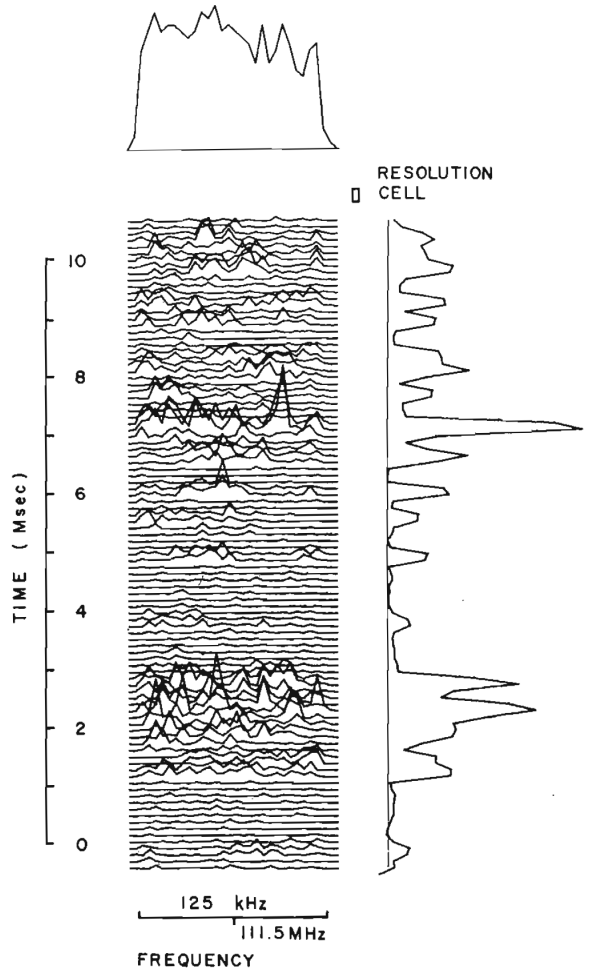


Figure: From Hawkins & Rickett (1975), "Pulsar Signal Processing", p. 108 via VDQI (page 134)

Data Density

$$\text{data density of a graphic} = \frac{\text{number of entries in data matrix}}{\text{area of data graphic}}$$

Maximize Data Density

- Depict more data
- Shrink the graphic
- Use multifunctioning graphical elements

Within reason!

Good data density?

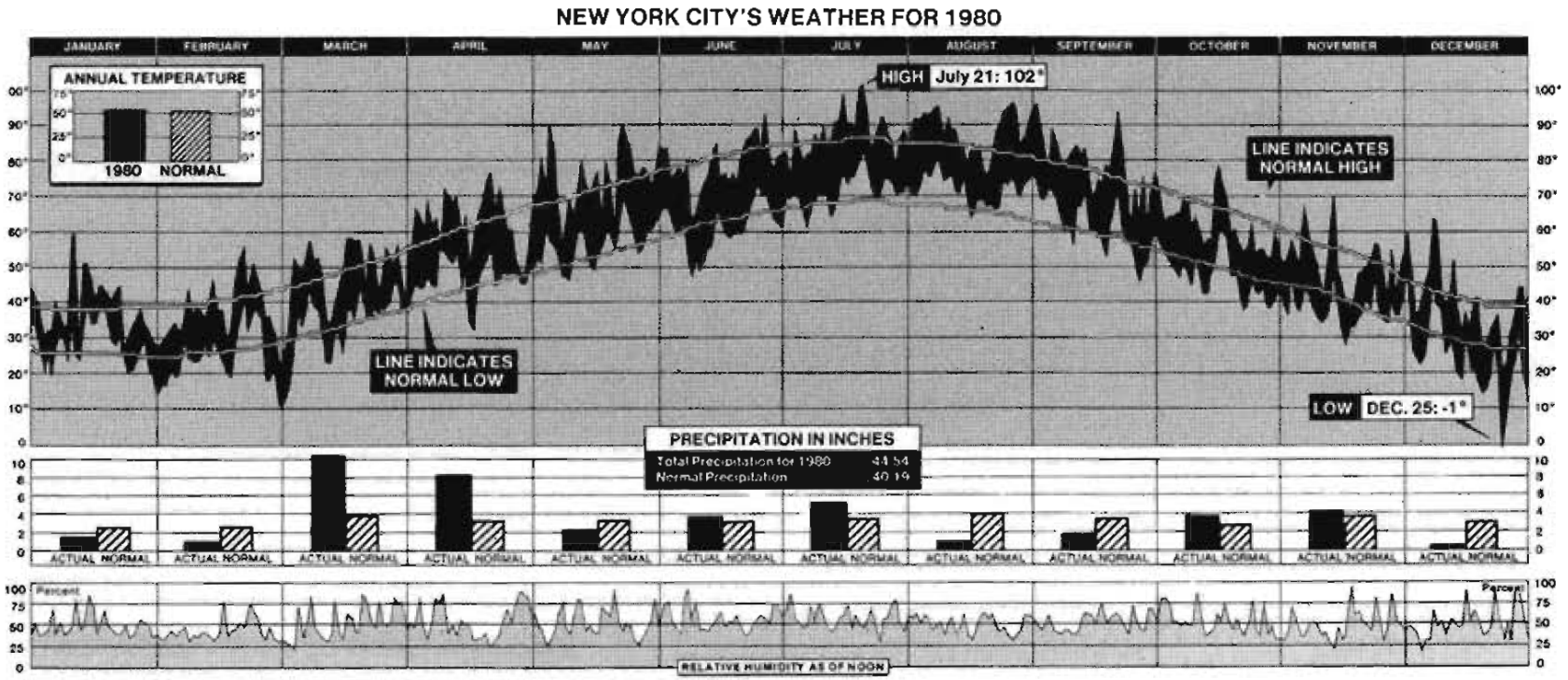


Figure: From ??? via VDQI (page 164)

Better data density?

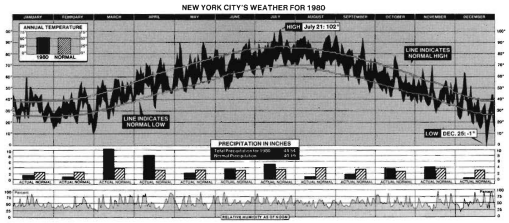


Figure: From ??? via VDQI (page 164)

Small Multiples

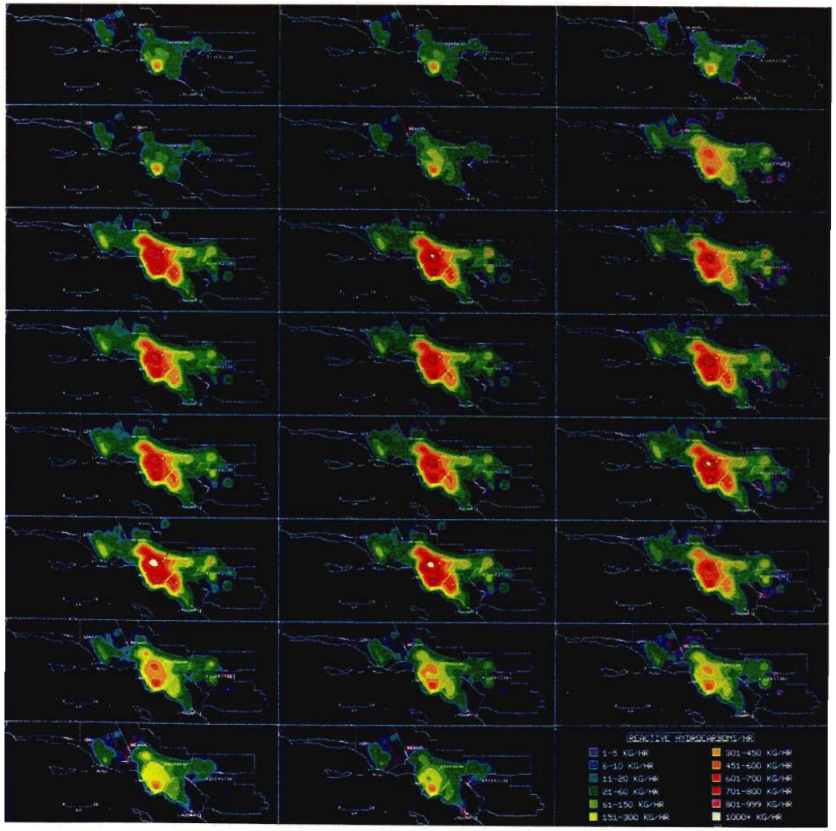


Figure: From video based on McRae, Goodin & Seinfeld (1982), "Development of a Second-Generation Mathematical Model for Urban Air Pollution" via VDQI (page 170)

Multifunctioning Graphical Elements

Advice

Mobilize every graphical element, perhaps several times over, to show the data.

Stem-and-leaf display

A stem-and-leaf display let's you show fairly detailed distribution information in the shape of a histogram.

Example (Data)

```
37, 33, 33, 32, 29, 28, 28, 23,
22, 22, 22, 21, 21, 21, 20, 20,
19, 19, 18, 18, 18, 18, 16, 15,
14, 14, 14, 12, 12, 9, 6
```

Example from **Lane @ OnlineStatBook**.

Example (S&L display 1)

```
3|2337
2|001112223889
1|2244456888899
0|69
```

Example (S&L display 2)

```
3|7
3|233
2|889
2|001112223
1|56888899
1|22444
0|69
```

Number Plots

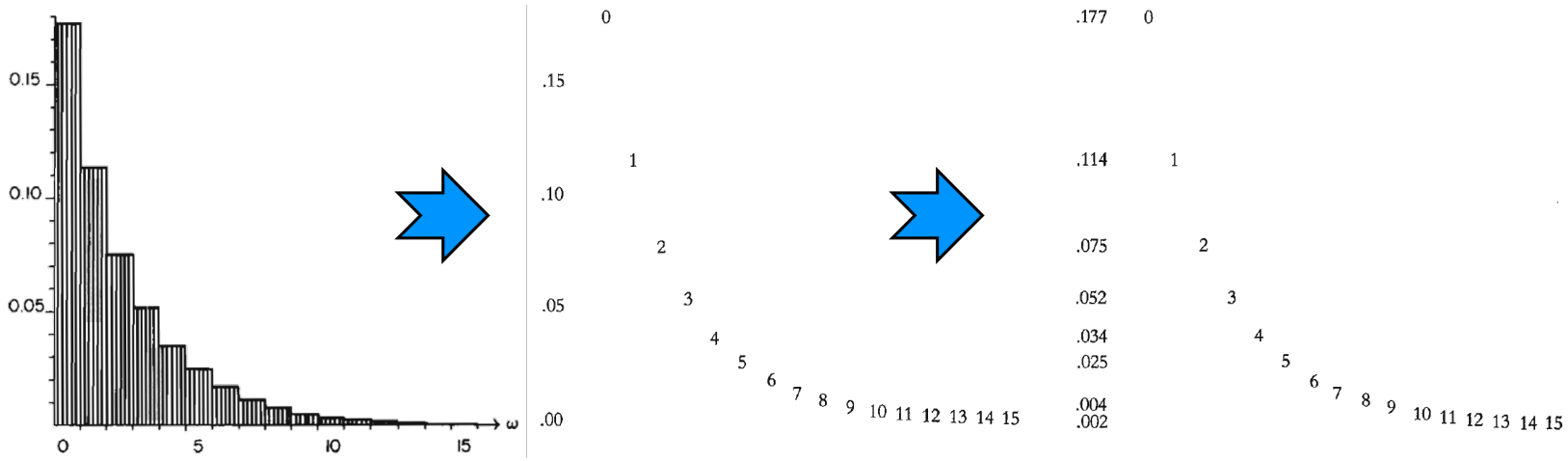


Figure: From stylesheet of the Journal of the American Statistical Association (left) and VDQI (page 150-151)

Quiver Plot

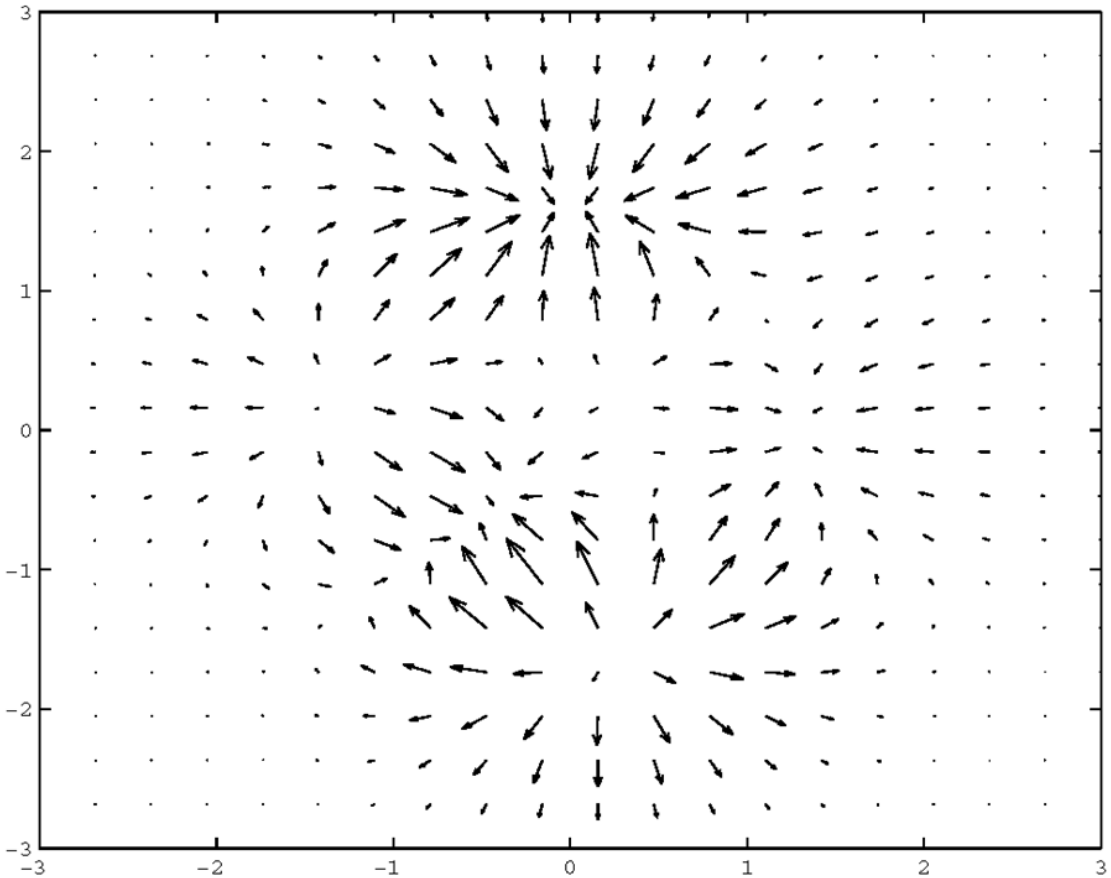


Figure: From **what-when-how**

Chernoff Faces

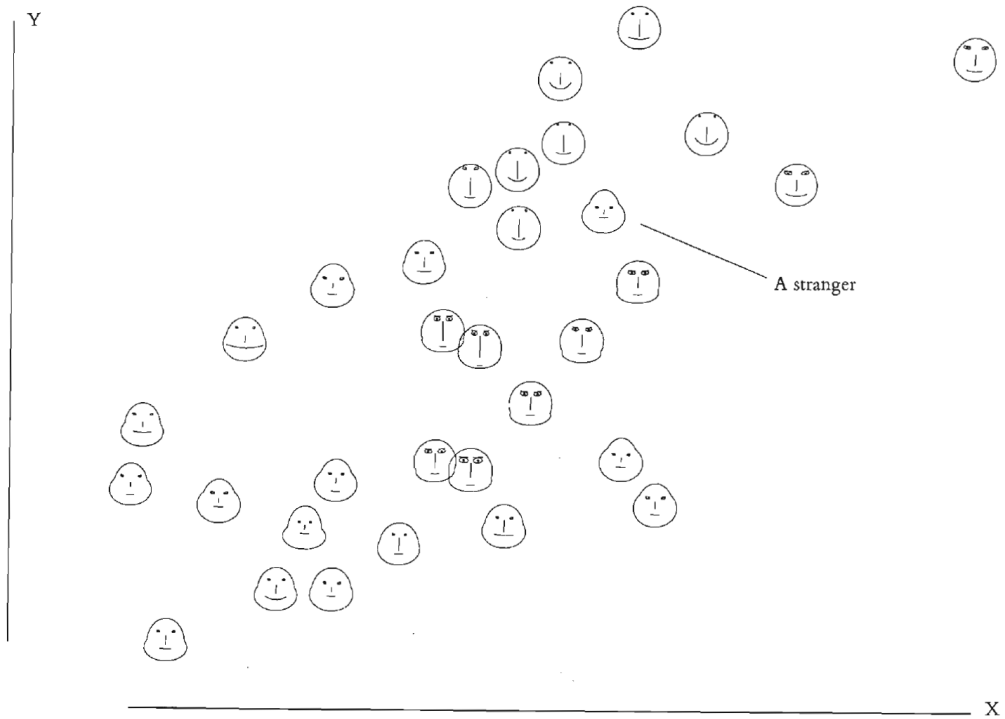


Figure: From Wainer & Thissen (1981), "Graphical Data Analysis" via VDQI (page 142)

See also Chernoff (1973), "The Use of Faces to Represent Points in k-Dimensional Space Graphically" and **Wikipedia**.

Meaningful marks

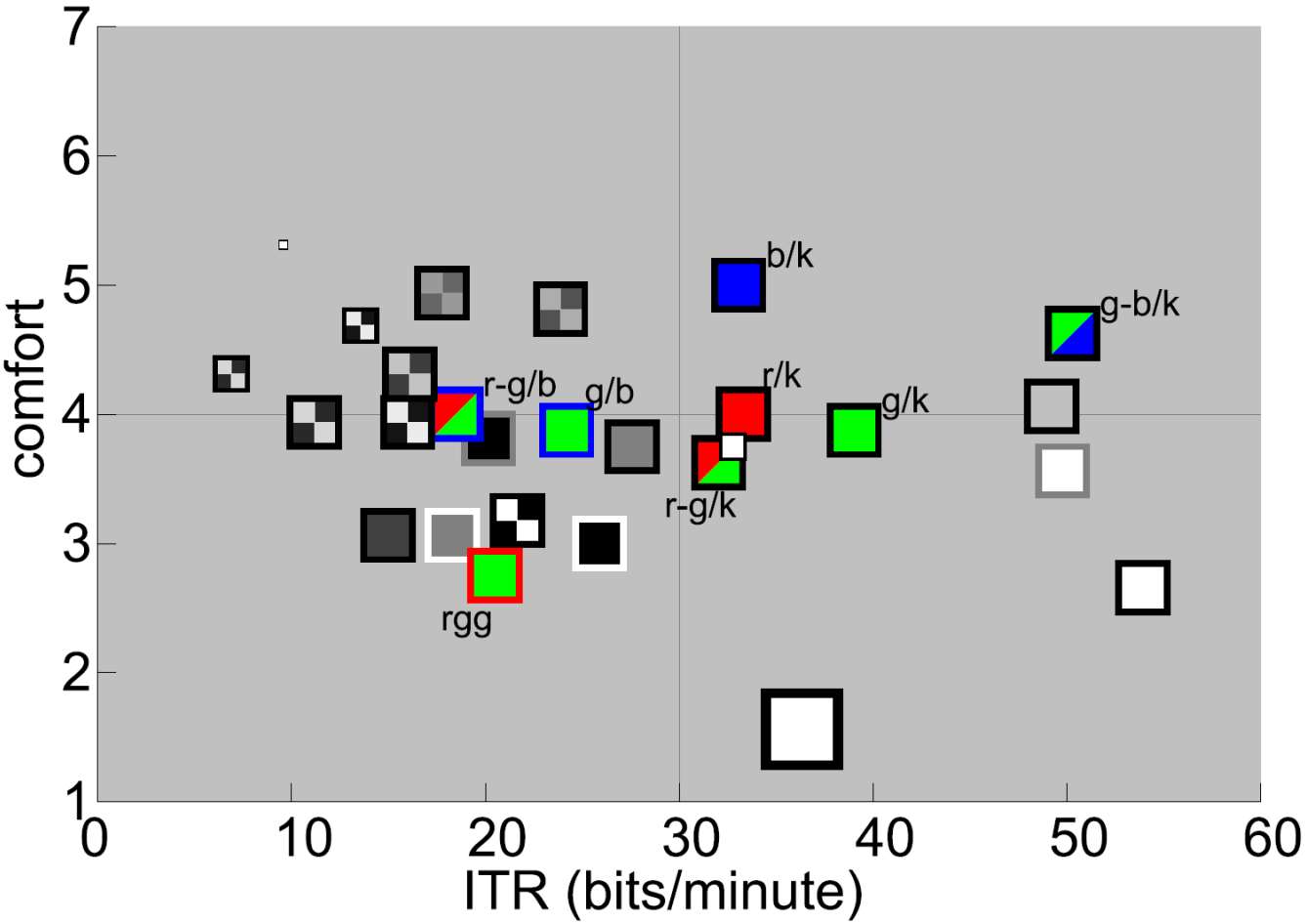


Figure: From **my master thesis** (PDF page 115)

Aesthetics

- Color
- Typography
- Proportions

Emphasize data

Figure: From VDQI (page 186)

Aspect ratio

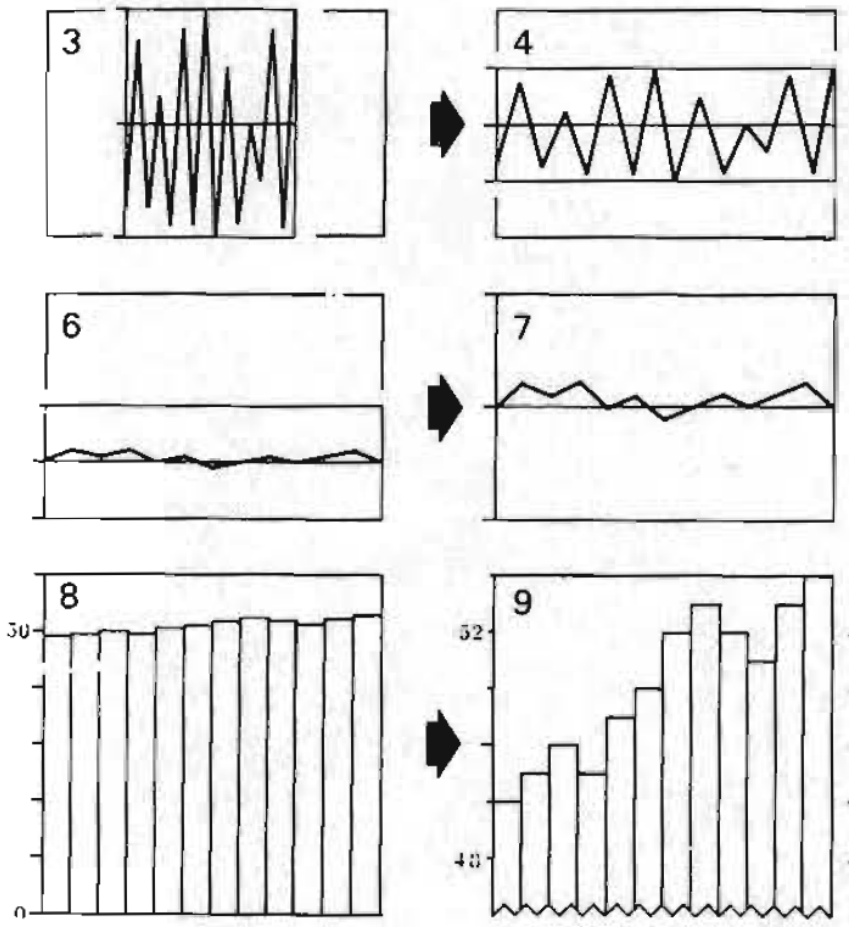


Figure: From Bertin (1973), "Semiologie Graphique" via VDQI (page 169)

Zoom in

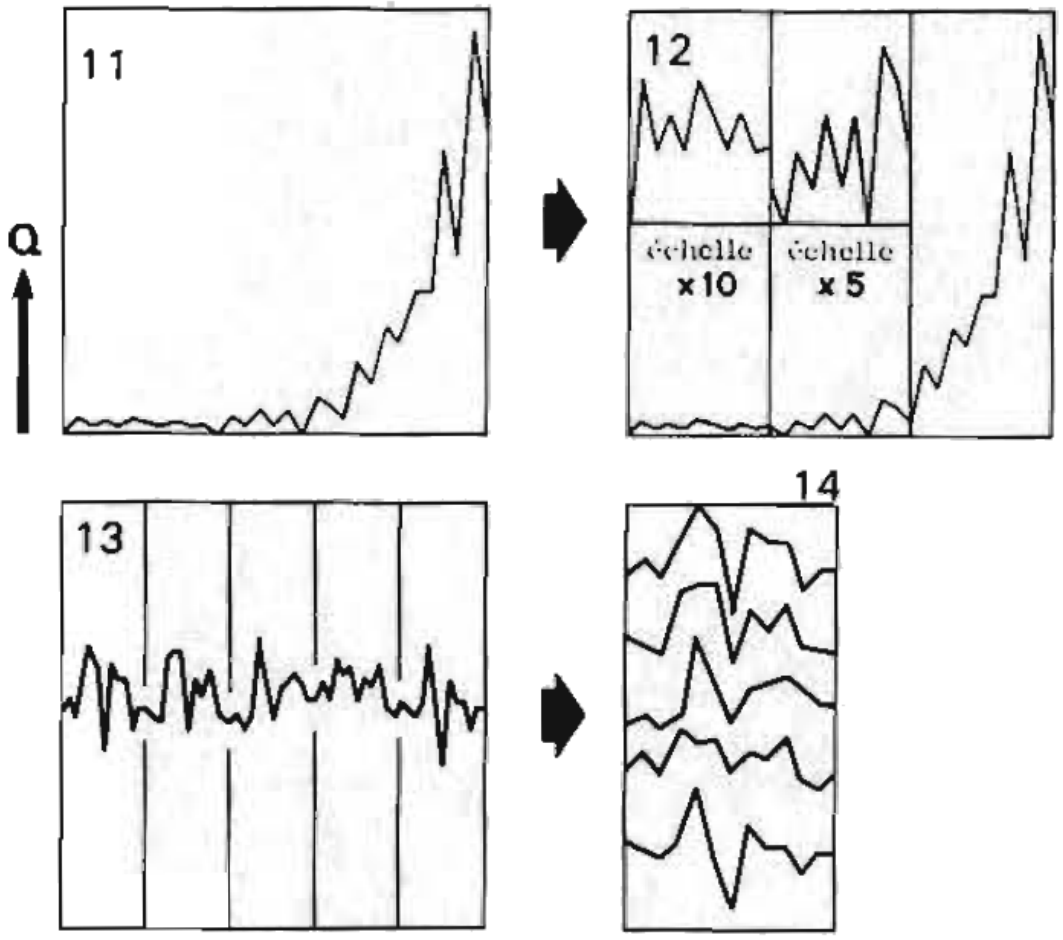


Figure: From Bertin (1973), "Semiologie Graphique" via VDQI (page 169)