

# Data Mining

COIS 4400H / AMOD 5400H

## Visualization Techniques (Chapter 3)

# Outline

Visualizing single attributes

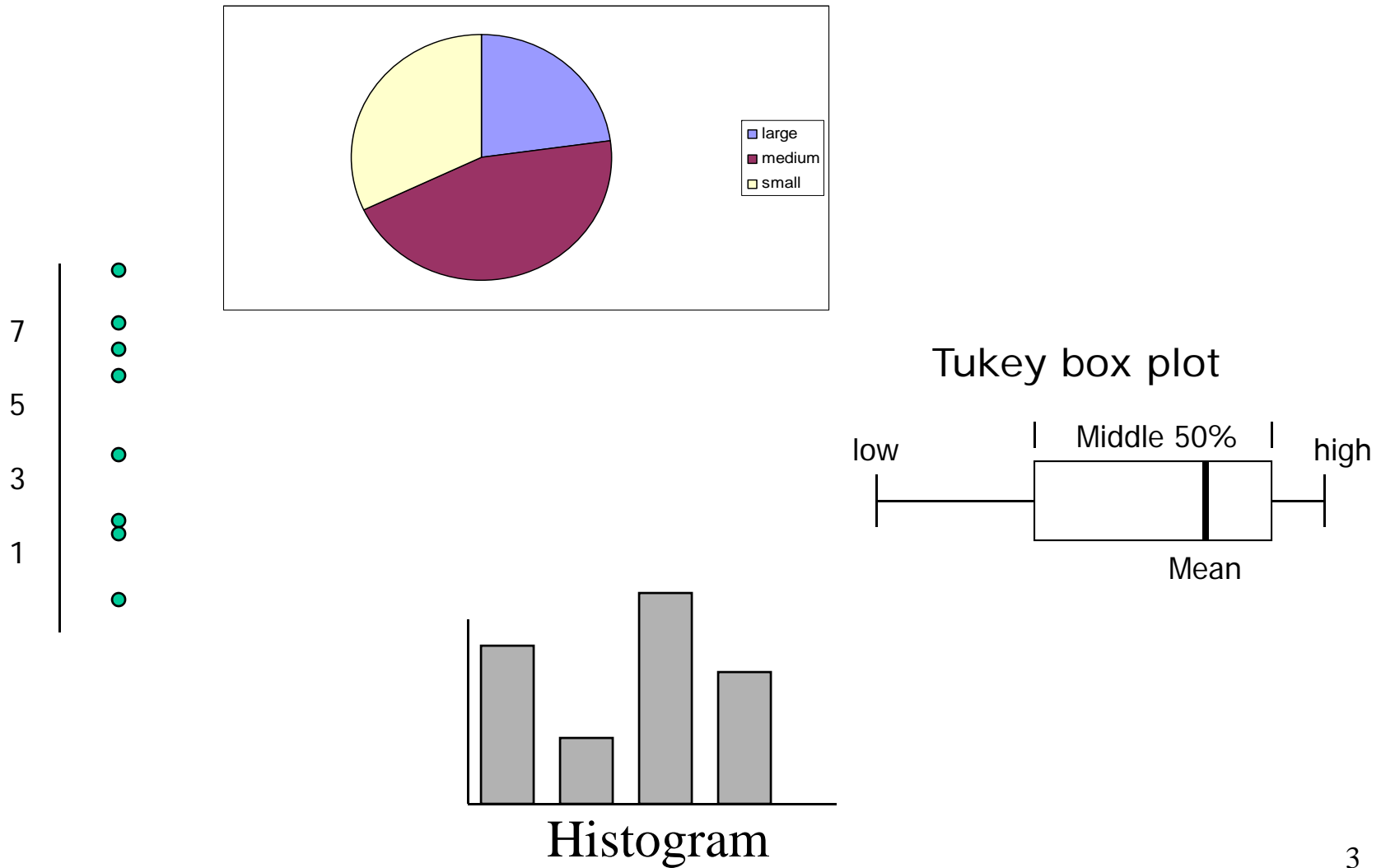
Visualizing a small number of attributes

Visualizing high-dimensional data

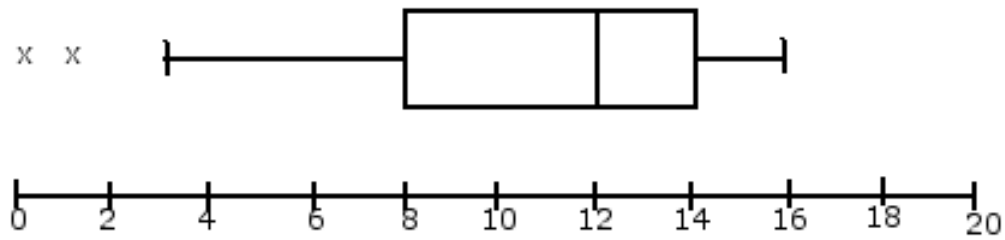
Examples

Tufte's Principles of Graphic Excellence

# 1-D (Univariate) Data



# Box plots



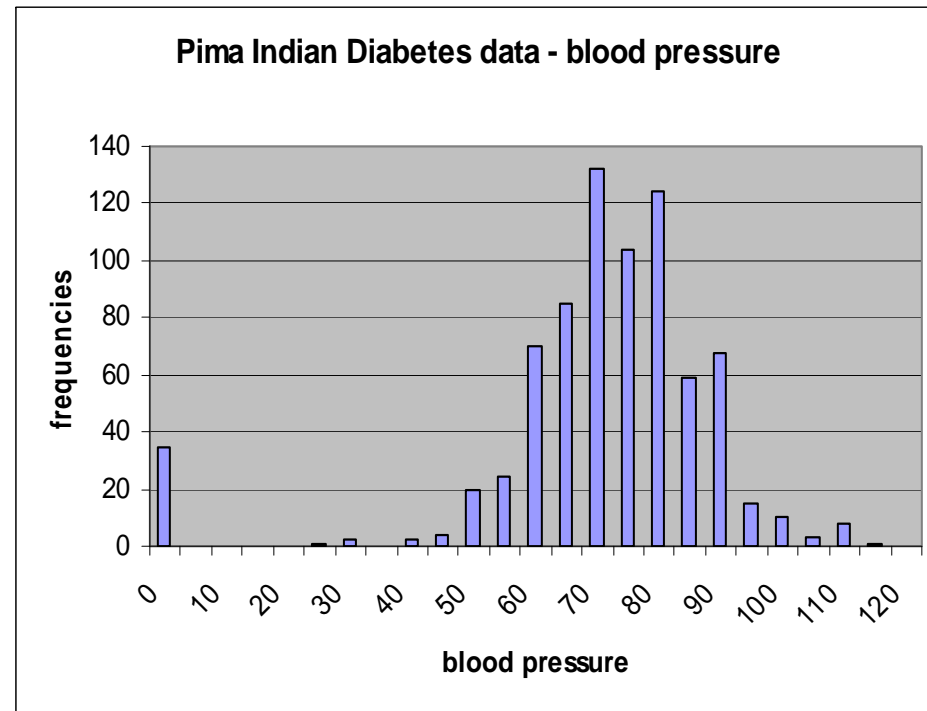
# Histogram

From National Institute of Diabetes and Digestive and Kidney Diseases: Pima Indian dataset

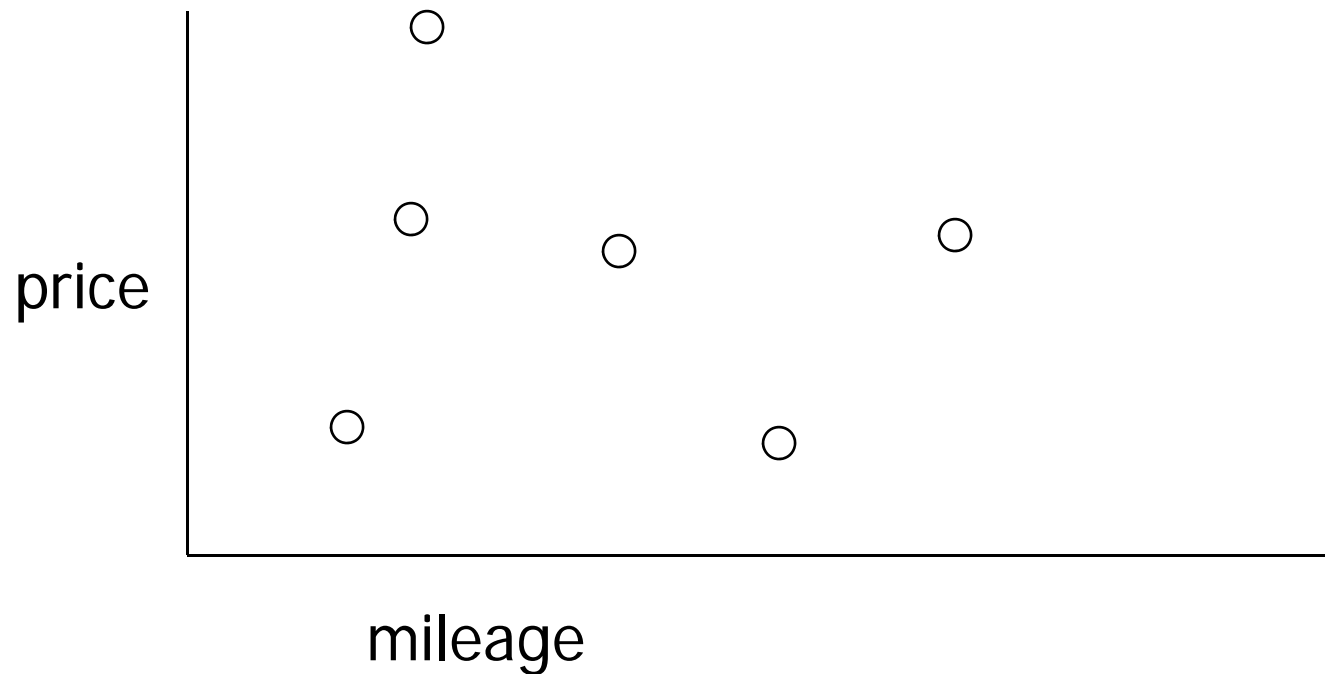
Binary classes (tested positive or negative for diabetes)

All 8 attributes are numeric-valued

768 instances

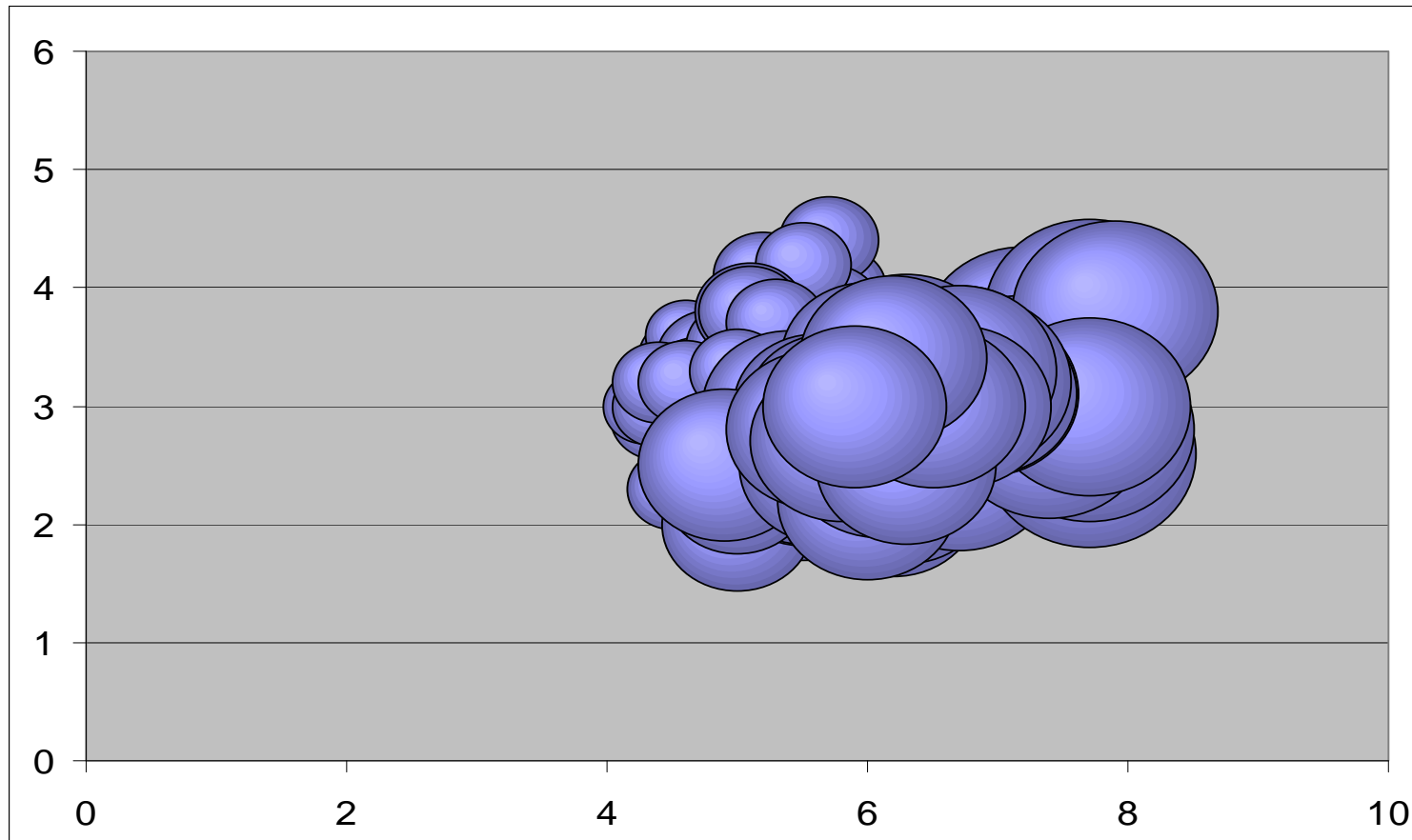


# 2-D (Bivariate) Data: Scatterplot

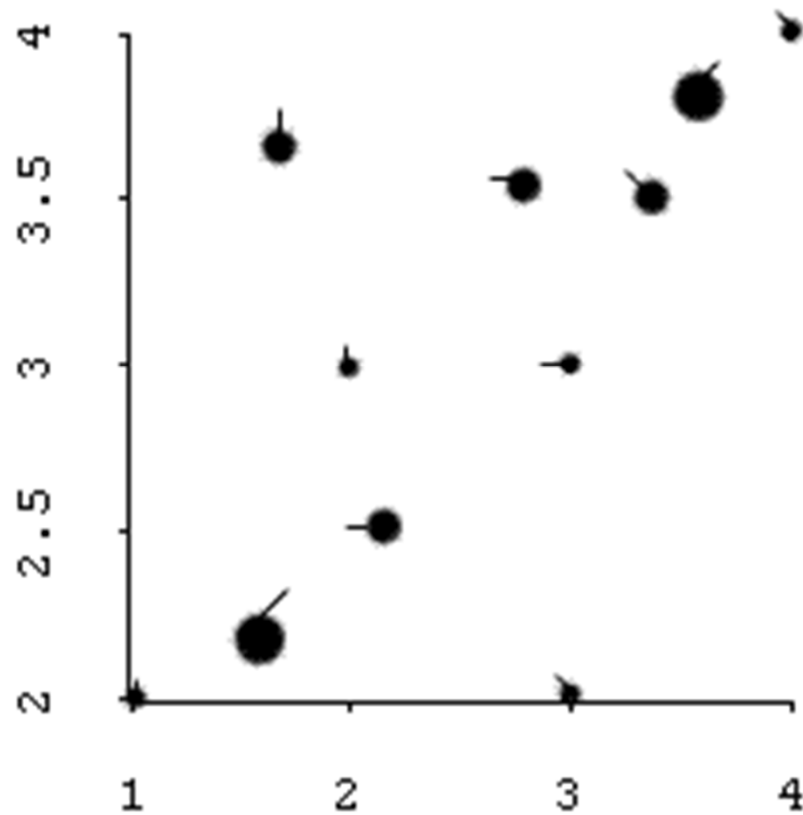


Shows relationship between pairs of variables

# Scatterplot with 3<sup>rd</sup> dimension (BubblePlot)

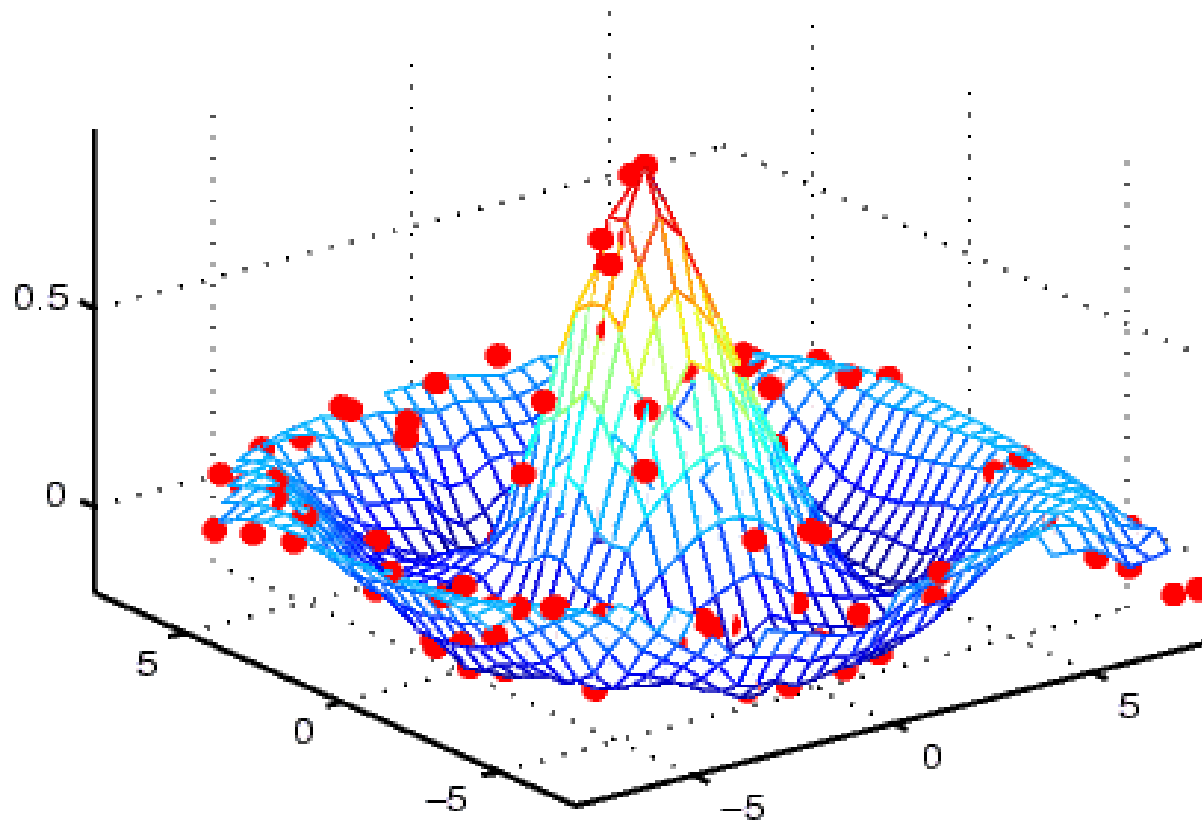


# Scatterplot with 5 dimensions





# 3-D Data (projection)



# Visualizing in 4+ Dimensions

Multiple Views

Scatterplot matrices

Trellis plots

Parallel Coordinates

Chernoff faces

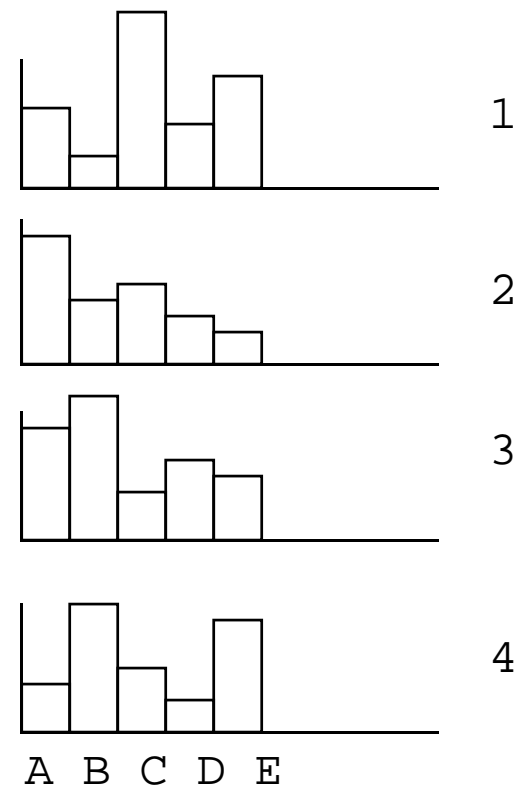
Star plots

Stick Figures

# Multiple Views

Give each variable its own display

	A	B	C	D	E
1	4	1	8	3	5
2	6	3	4	2	1
3	5	7	2	4	3
4	2	6	3	1	5



# Example: Visualizing Iris Data



Iris setosa

sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
...	...	...	...
5.9	3	5.1	1.8

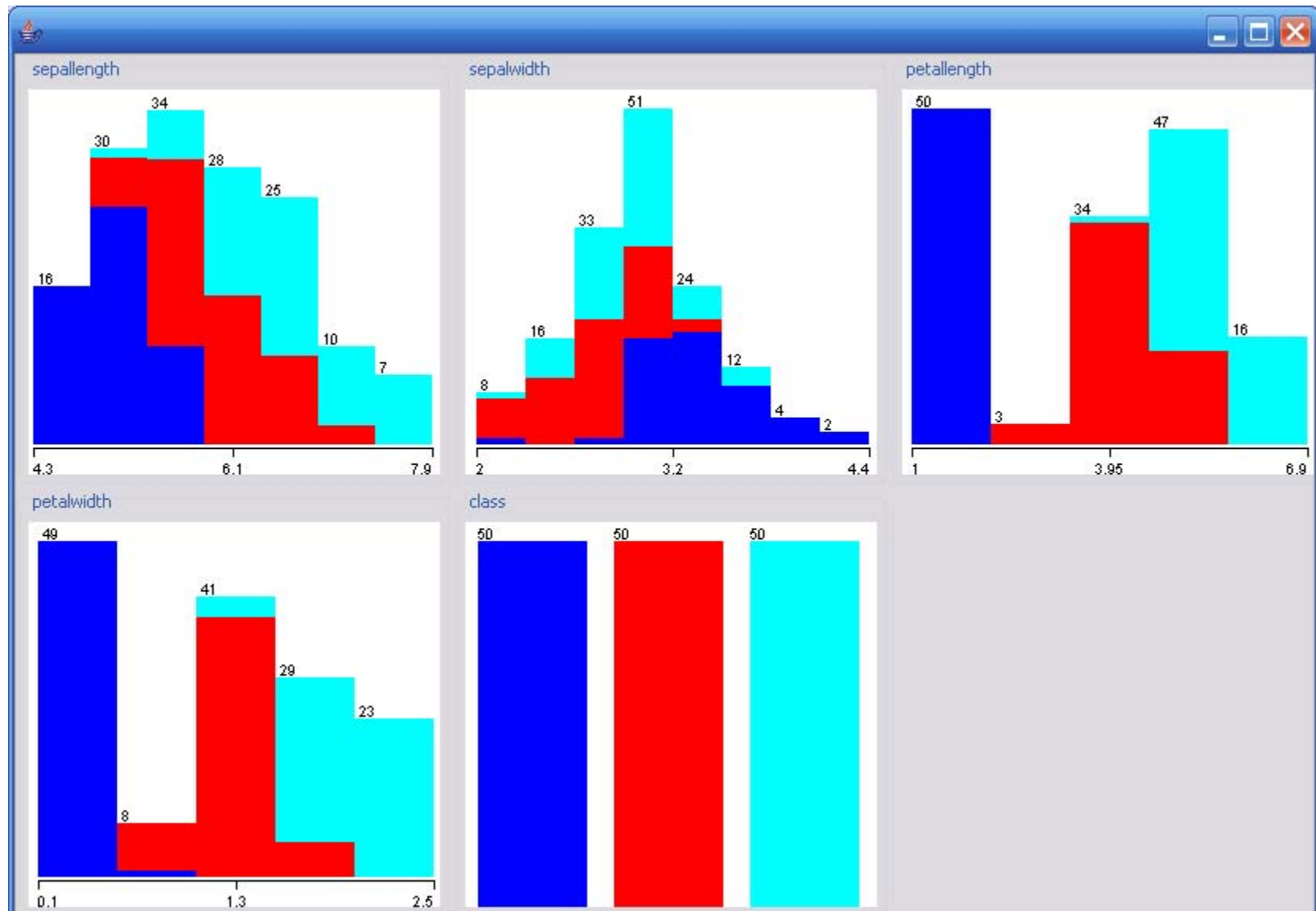


Iris versicolor



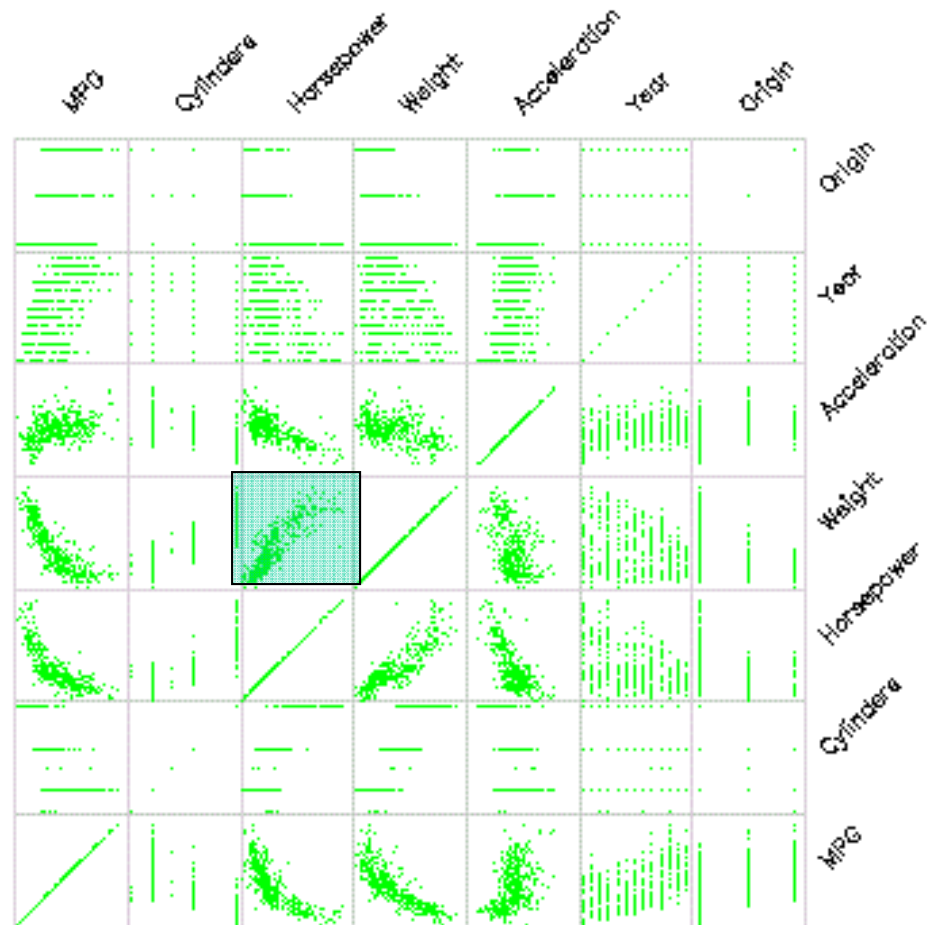
Iris virginica

# Individual Attributes for Iris Data

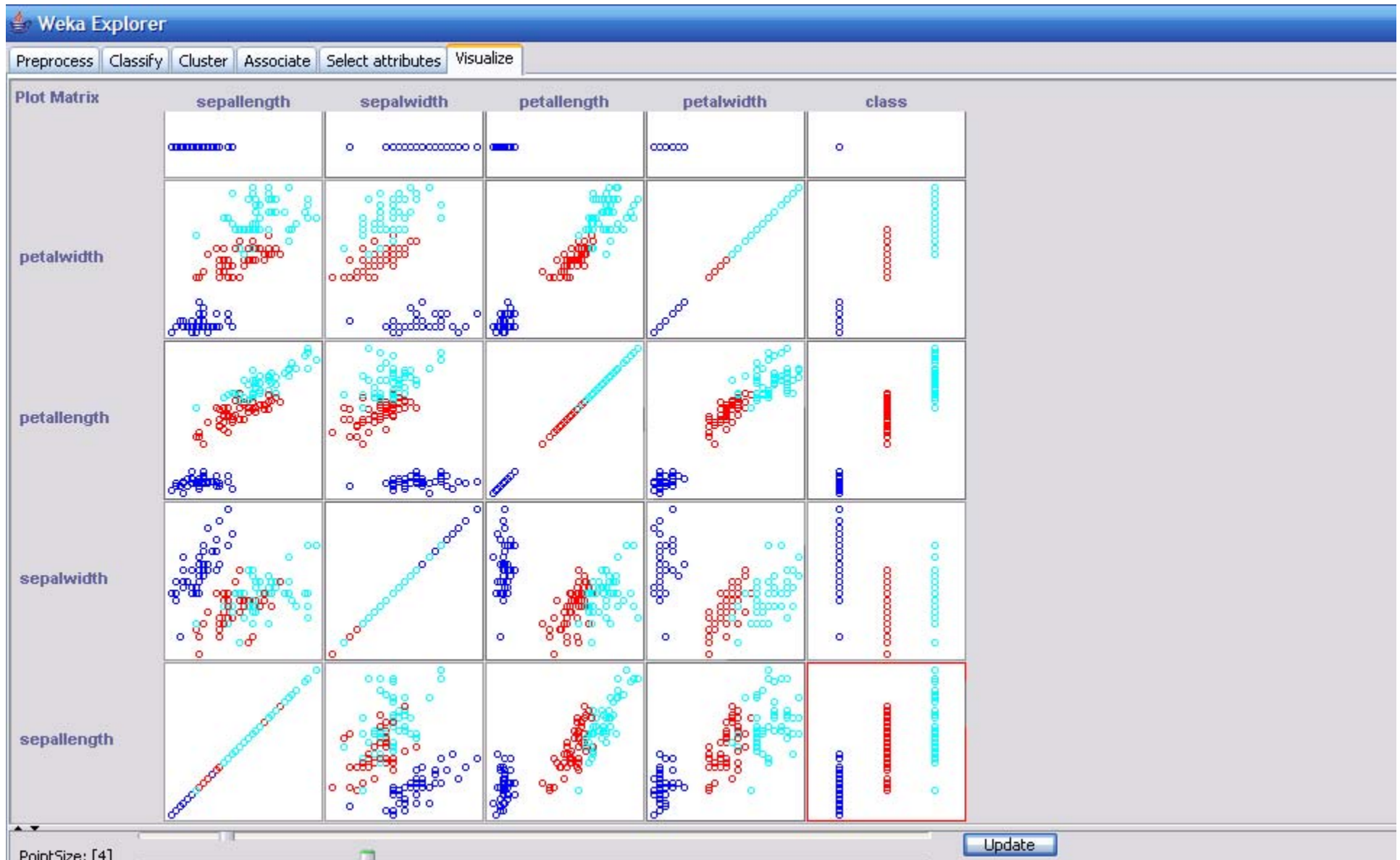


# Scatterplot Matrix

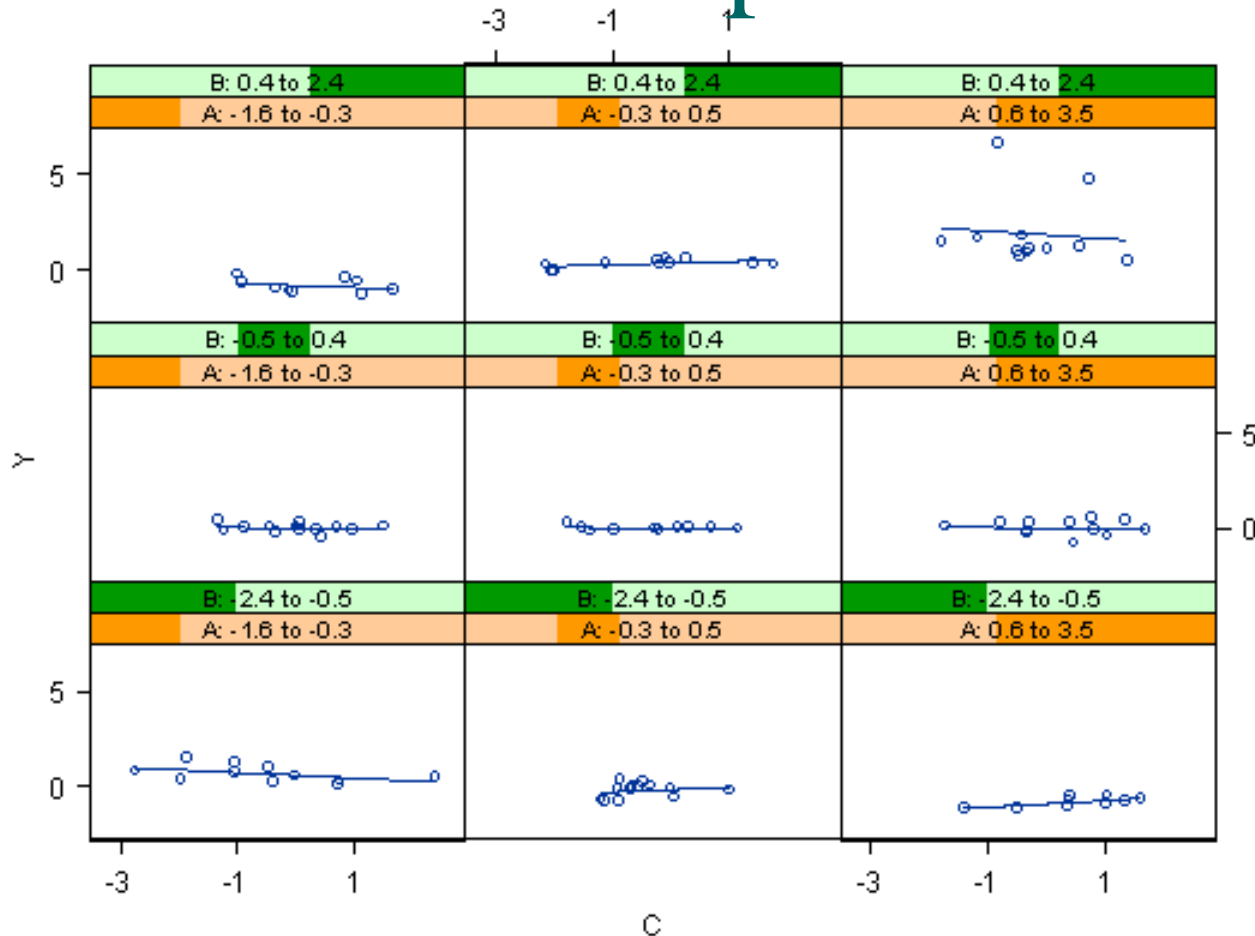
Represent each possible pair of variables in their own 2-D scatterplot



# Scatterplot Matrix for Iris Data



# Trellis plots

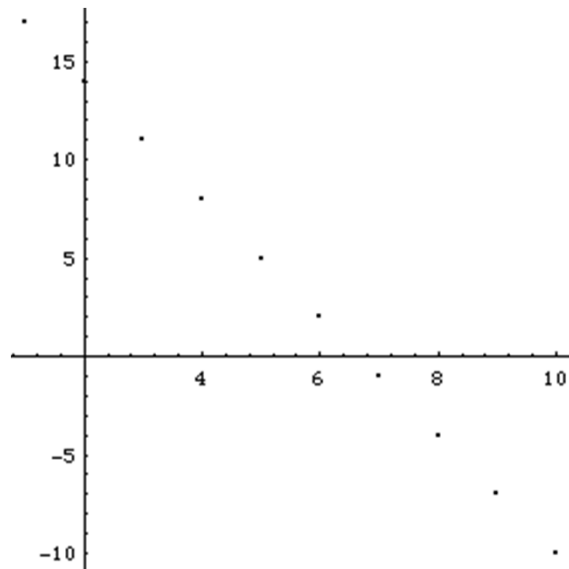


- uses multiple bivariate plots
- a pair of variables is fixed, plots are then produced based on the values of one or more of the other variables
- trellis plots are like mega-plots: they can be produced with any kind of component graph (box plots, histograms, contour plots....)

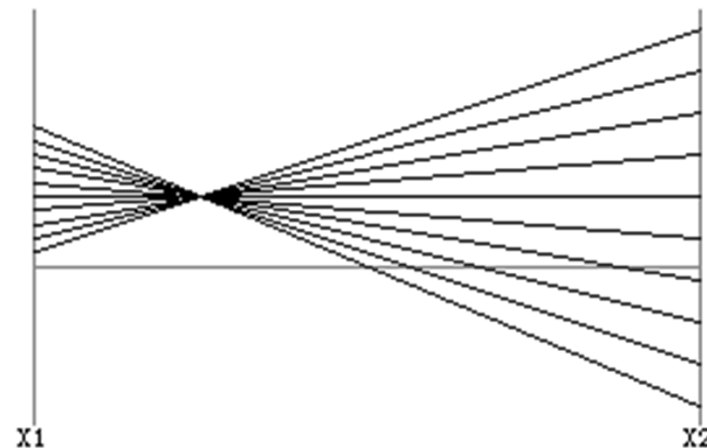


# Parallel Coordinates

- Encode variables along a horizontal row
- Vertical line specifies values

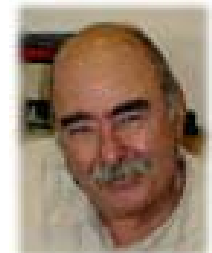


Dataset in Cartesian coordinates



Same dataset in parallel coordinates


Invented by  
Alfred Inselberg  
while at IBM, 1985



# Parallel Coordinates

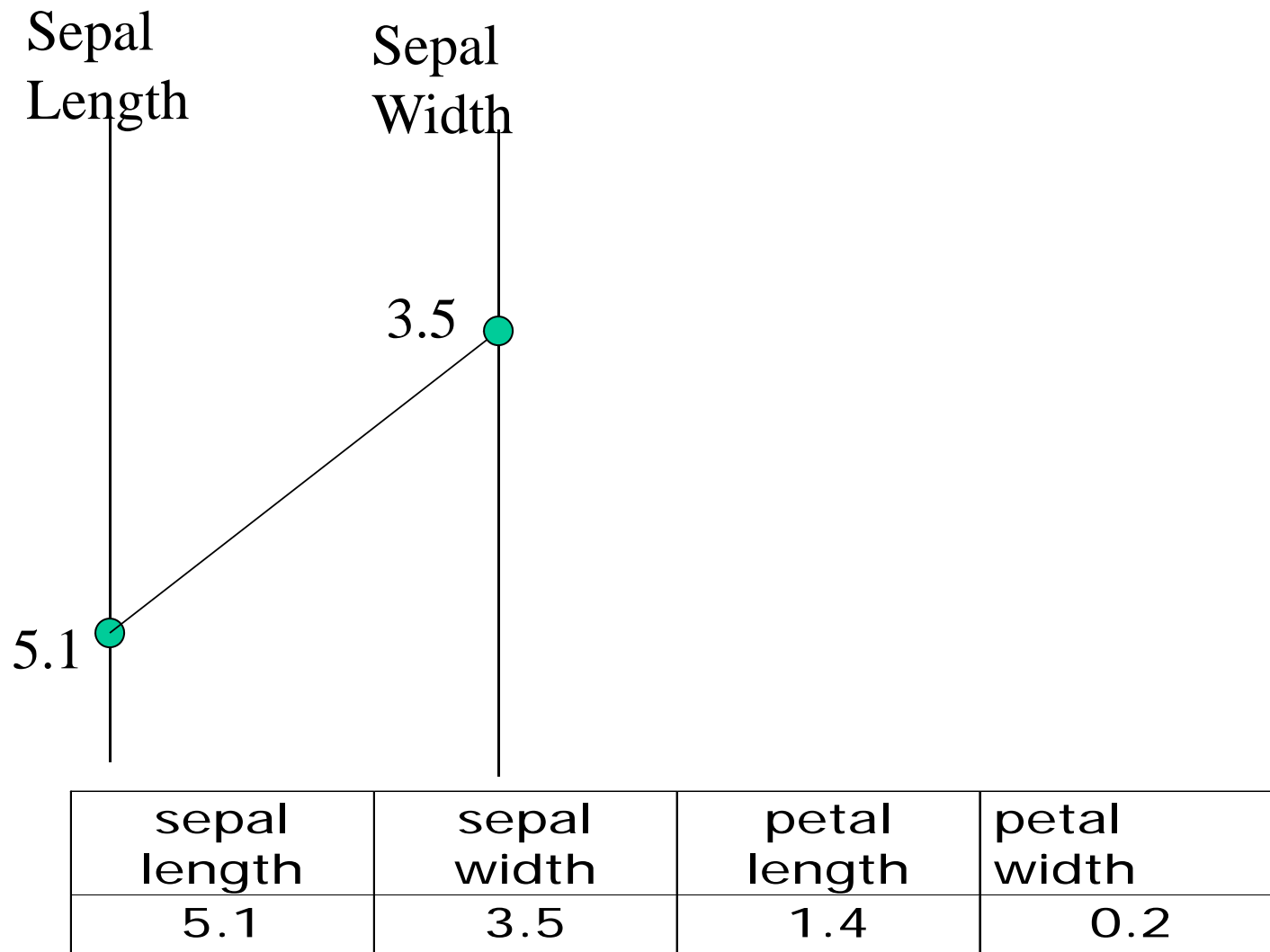
Sepal  
Length

5.1

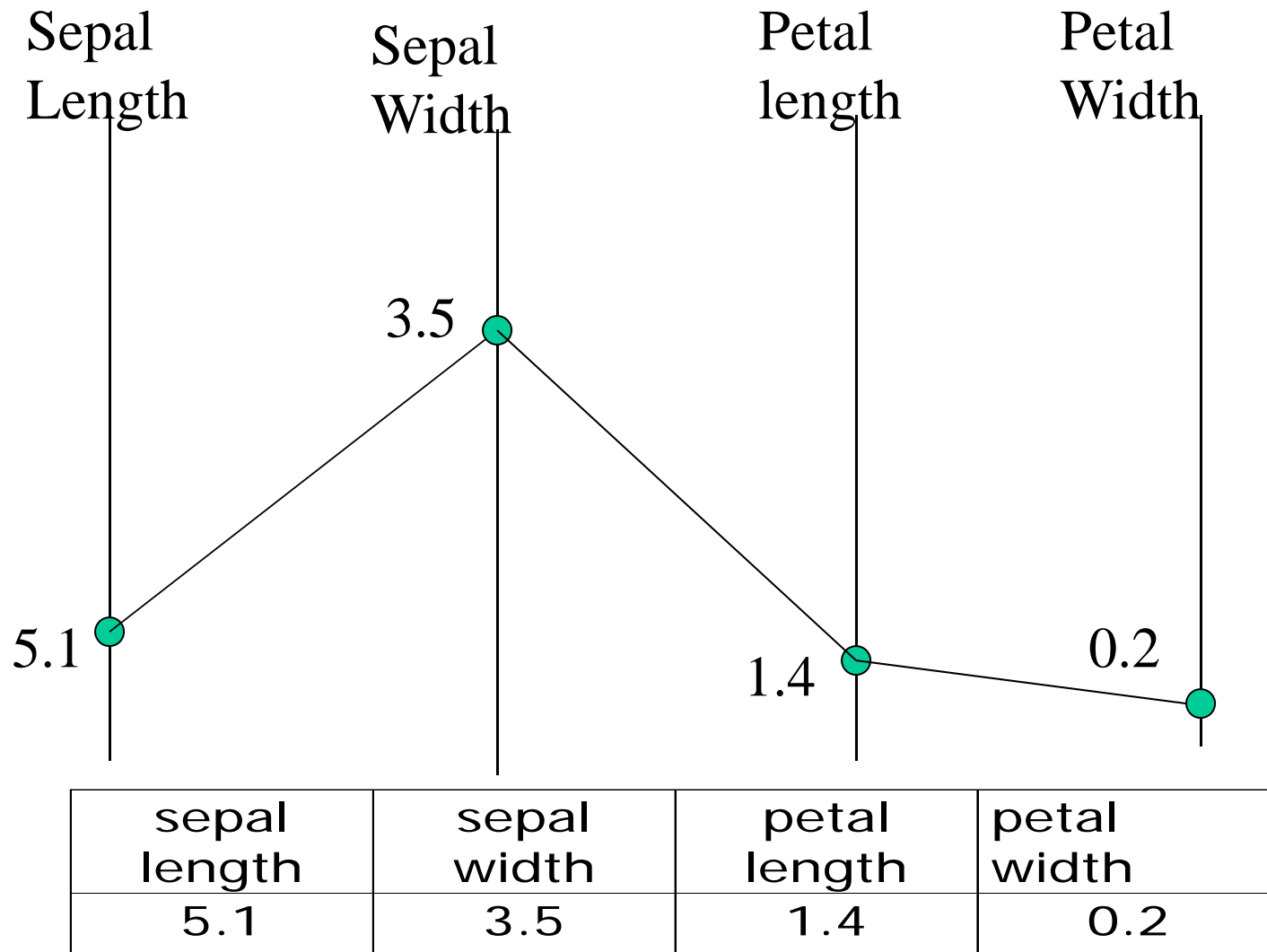


sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

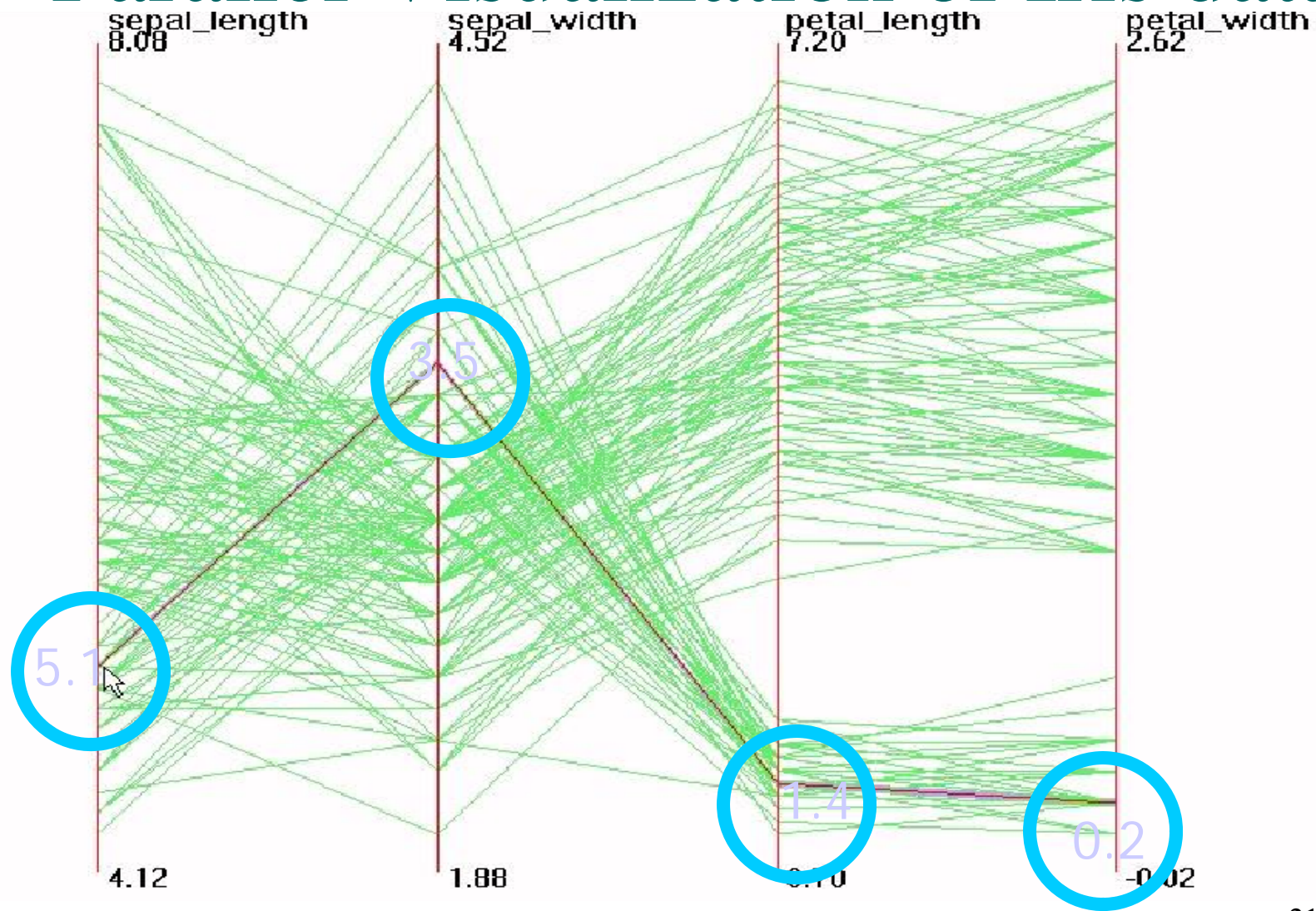
# Parallel Coordinates: 2 D



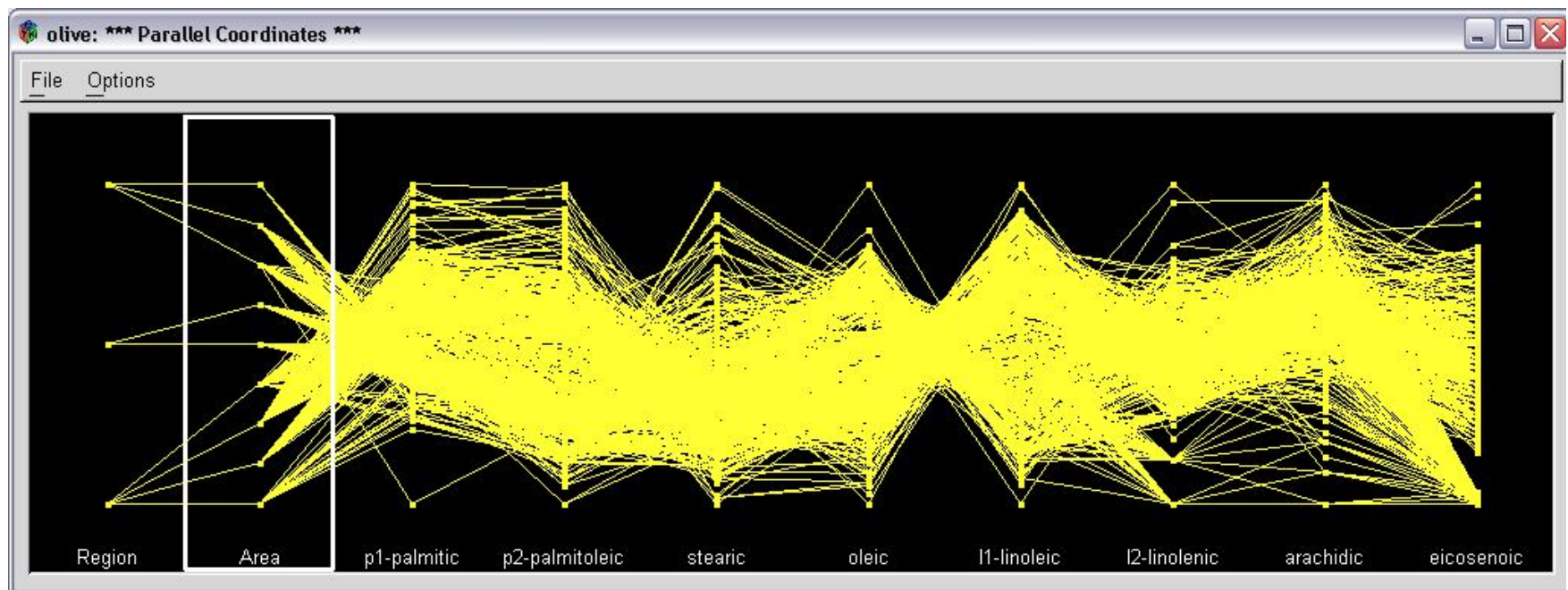
# Parallel Coordinates: 4 D



# Parallel Visualization of Iris data

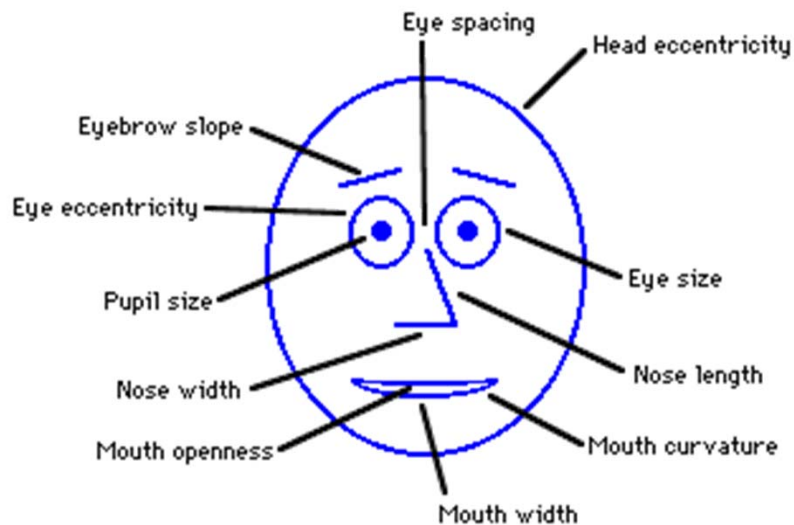


# Parallel coordinates: Olive data



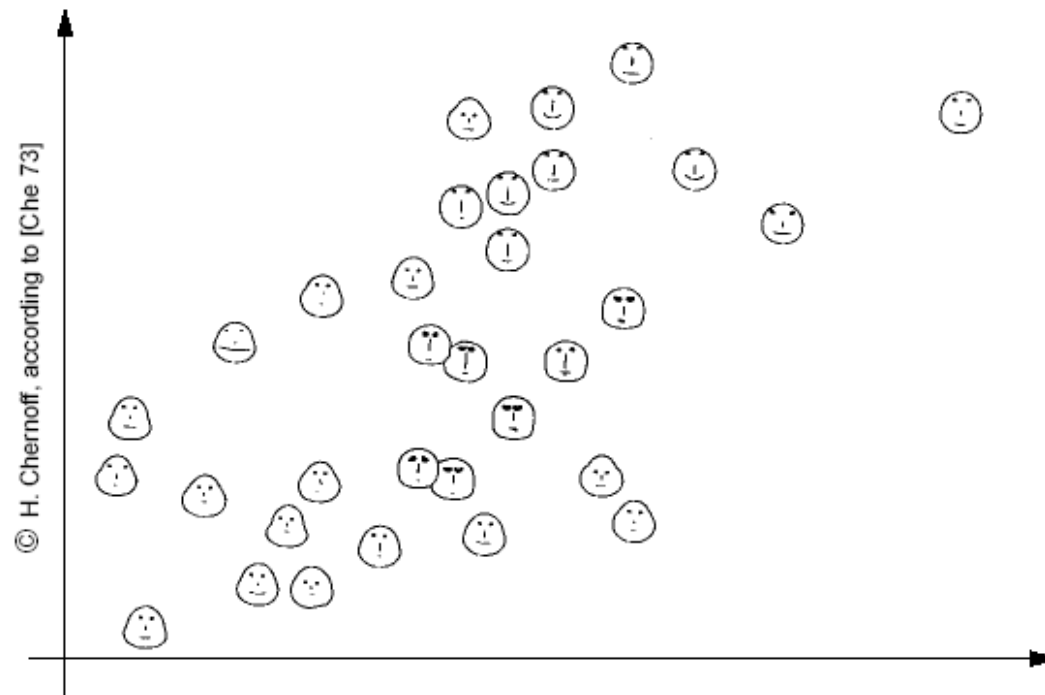
# Chernoff Faces

Encode values of different variables in characteristics of human face



Cute applets: <http://www.cs.uchicago.edu/~wiseman/chernoff/>  
<http://hesketh.com/schampeon/projects/Faces/chernoff.html>

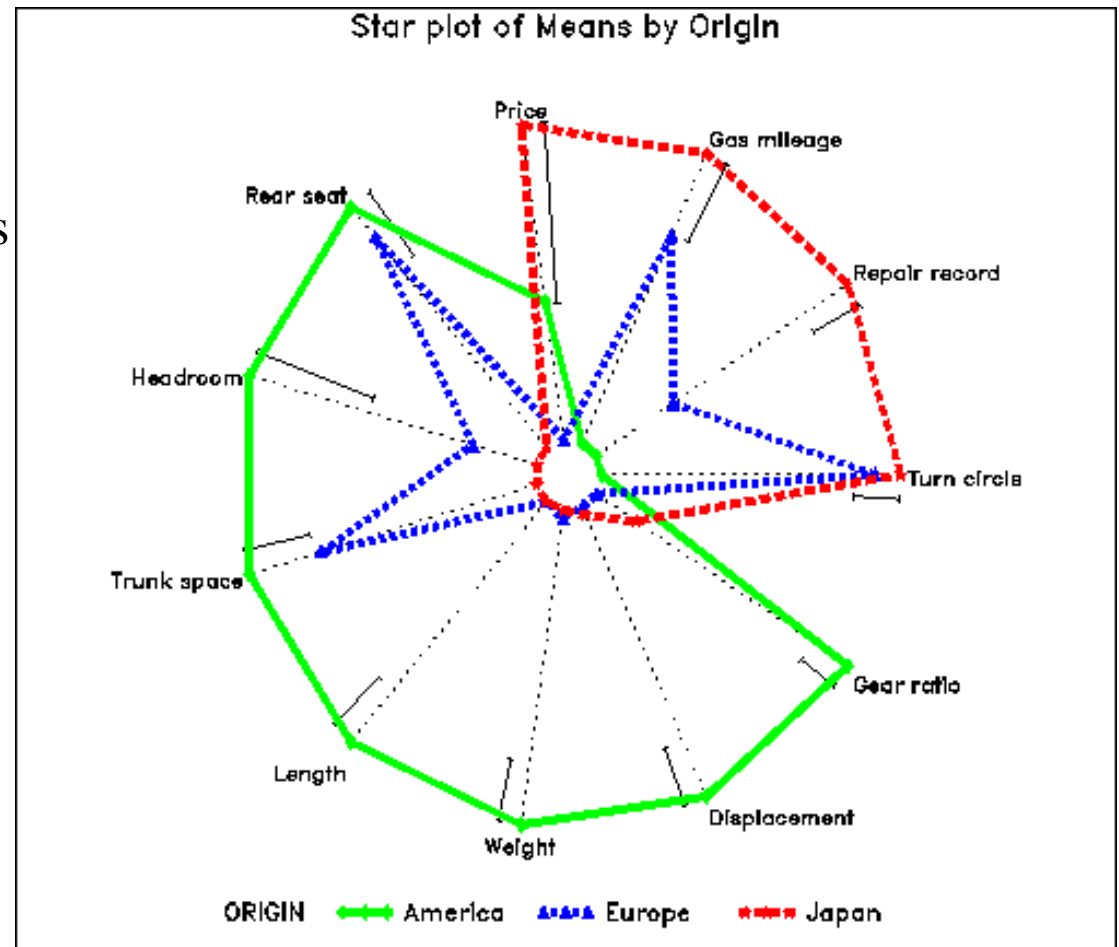
# Chernoff faces, example





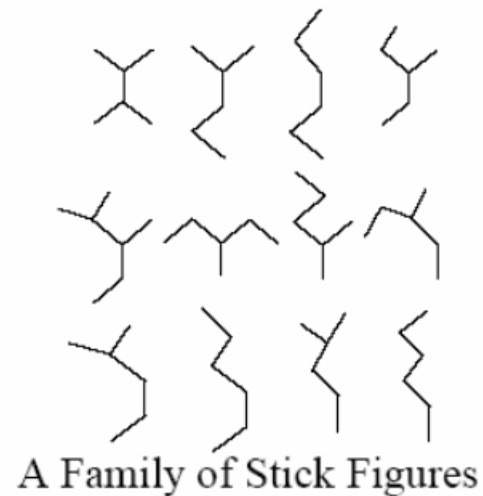
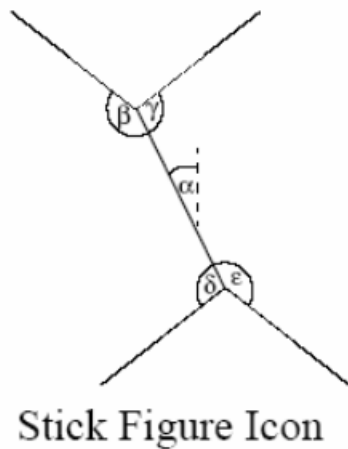
# Star plots

- spokes represent variables
- length of spoke corresponds to value of variable



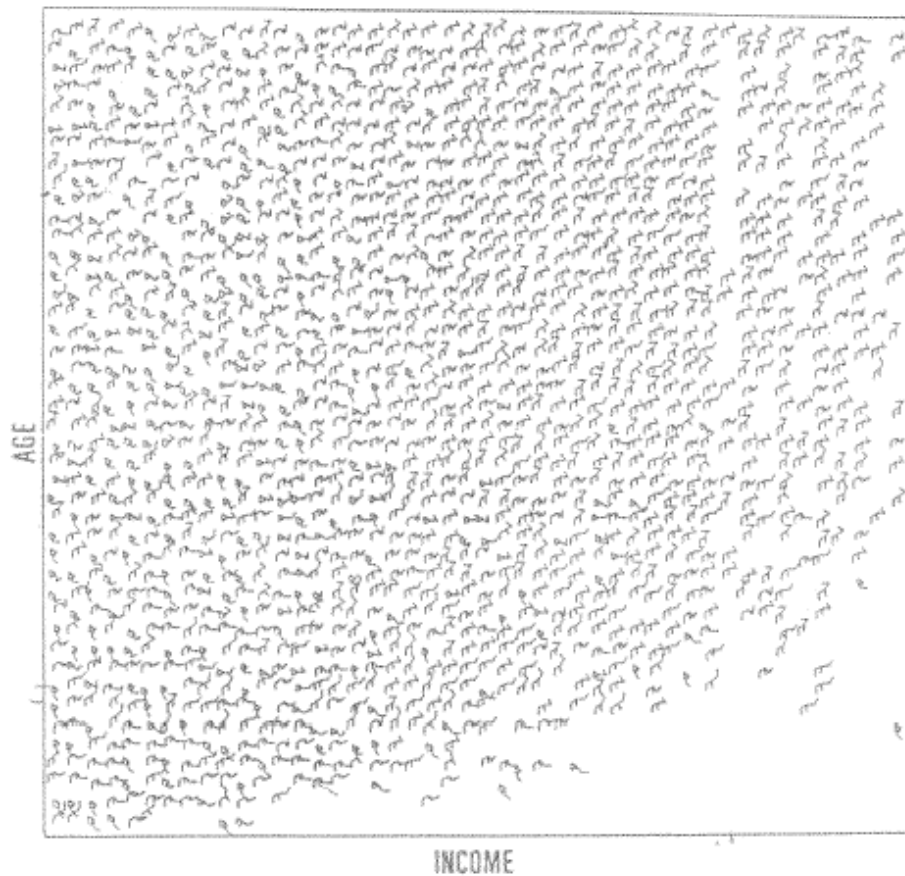
# Stick Figures

- Two variables are mapped to X, Y axes
- Other variables are mapped to limb lengths and angles
- Texture patterns can show data characteristics



# Stick figures

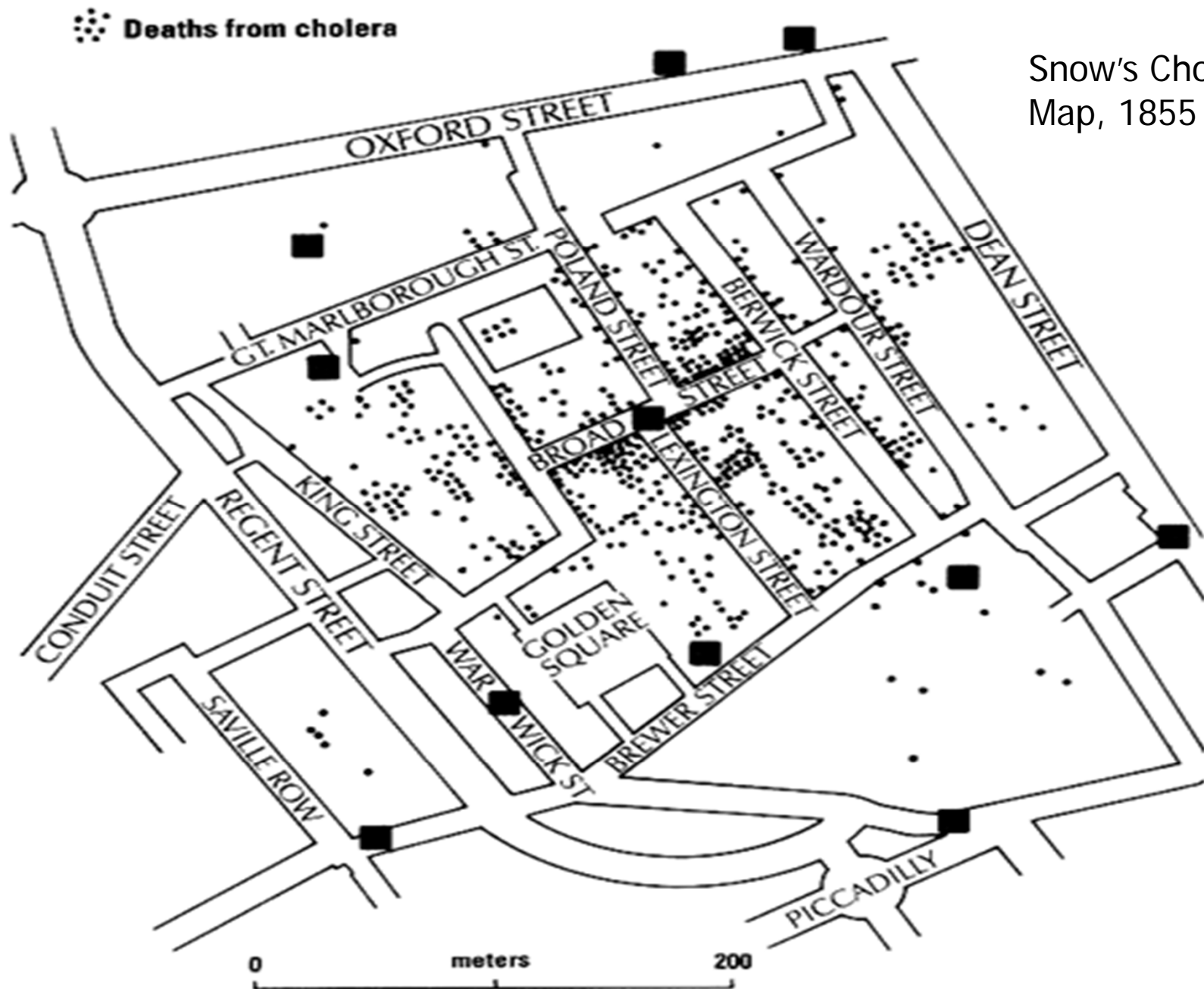
census data showing  
age, income, sex,  
education, etc.



used by permission of G. Grinstein, University of Massachusetts at Lowell

Closed figures  
correspond to women  
and we can see more  
of them on the left.

Note also a young  
woman with high  
income

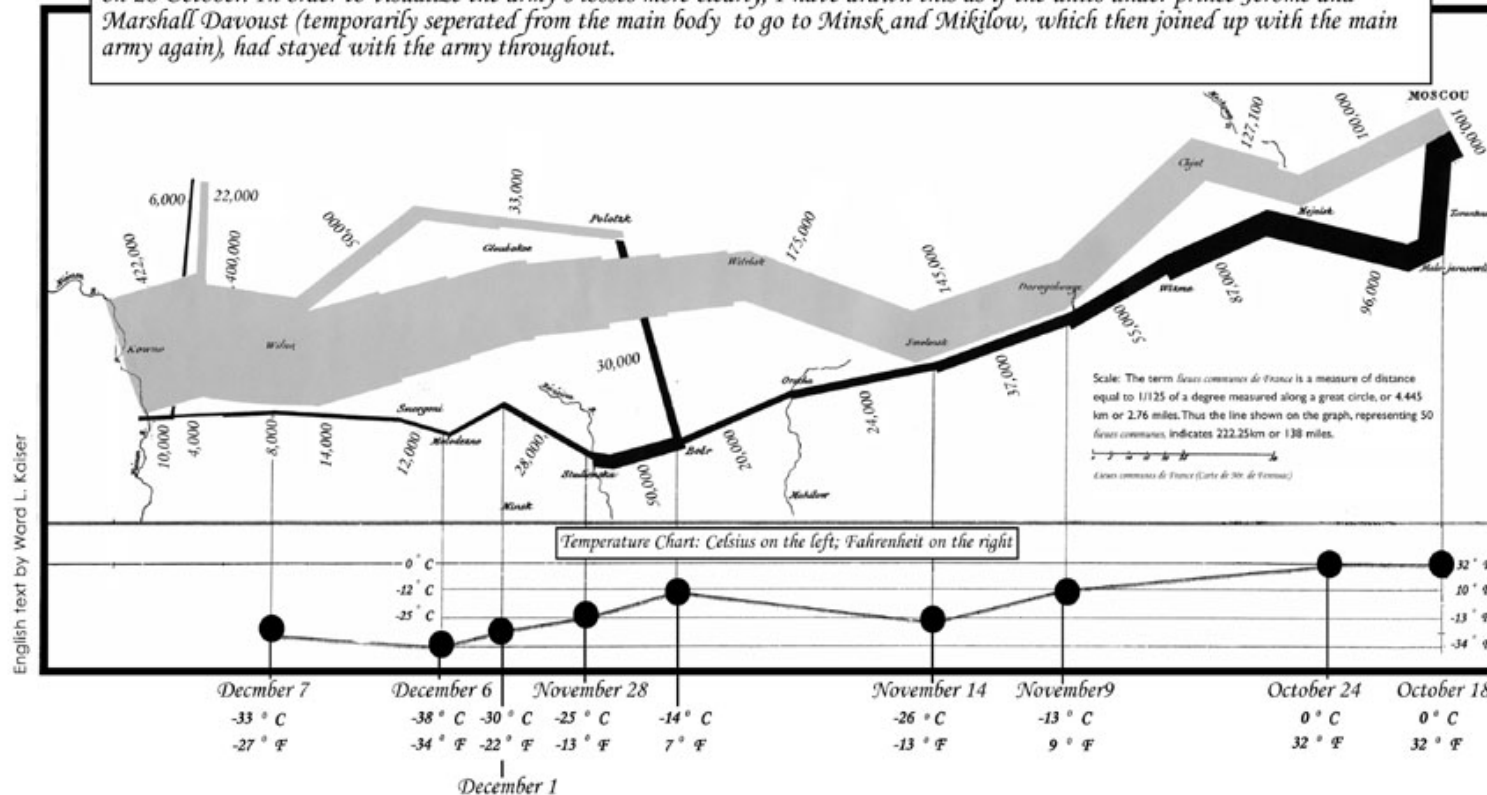


(E. Tufte, "The Visual Display of Quantitative Information", pg. 24)

Map representing the losses over time of French army troops during the Russian campaign, 1812-1813.  
Constructed by Charles Joseph Minard, Inspector General of Public Works retired.

Paris, 20 November 1869

The number of men present at any given time is represented by the width of the grey line; one mm. indicates ten thousand men. Figures are also written besides the lines. Grey designates men moving into Russia; black, for those leaving. Sources for the data are the works of messrs. Thiers, Segur, Fezensac, Chambray and the unpublished diary of Jacob, who became an Army Pharmacist on 28 October. In order to visualize the army's losses more clearly, I have drawn this as if the units under prince Jerome and Marshall Davoust (temporarily separated from the main body to go to Minsk and Mikilow, which then joined up with the main army again), had stayed with the army throughout.

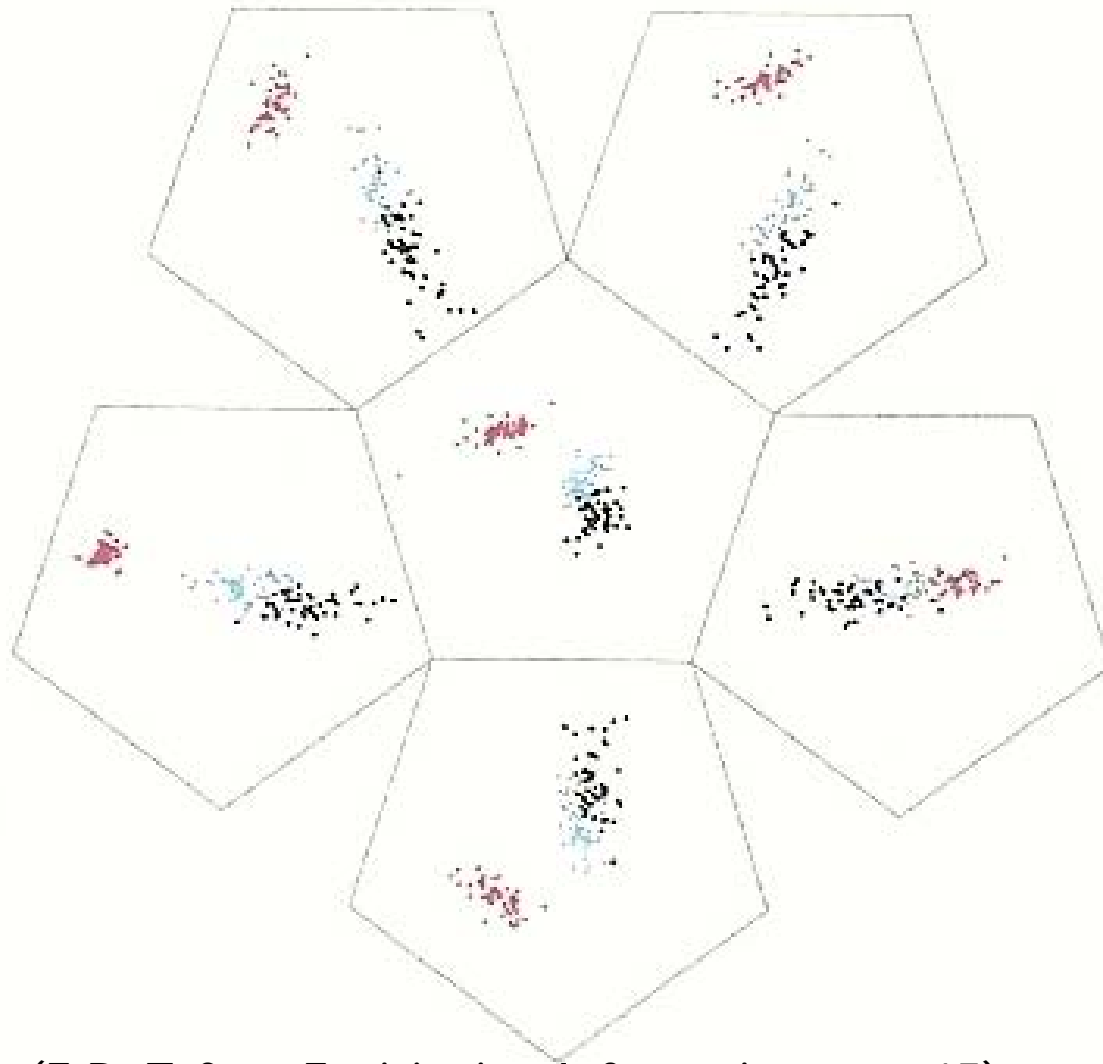


Editor's note: dates & temperatures are only referenced for the retreat from Moscow  
© 2001, ODT Inc. All rights reserved.

Figure 58. Minard's map of Napoleon's Russian campaign.

This graphic has been translated from French to English and modified to most effectively display the temperature data.

# Alternative view of Iris Data



(E.R. Tufte, "Envisioning Information", pg. 15)



		日	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
2月	54年	昼	☂	☂	☀	☂	☀	☂	☂	☂	☂	☂	☀	☀	☀	☀	☂	☀	☀	☂	☂	☀	☀	☀	☂	☂	☀	☂	☂	☂	
	夜	☂	☂	☂	★	☂	★	★	★	★	★	☂	★	★	★	☂	★	☂	☂	☂	★	☂	★	☂	☂	★	☂	☂	☂	☂	
	55年	昼	☂	☂	☂	☀	☂	☂	☂	☂	☀	☀	☀	☂	☀	☀	☂	☂	☂	☂	☀	☀	☂	☂	☂	☂	☂	☀	☀	☀	☀
	夜	☂	☂	★	☂	☂	☂	★	★	☂	★	★	★	☂	★	★	★	☂	☂	★	★	★	☂	☂	★	☂	★	☂	★	★	★
	56年	昼	☂	☀	☀	☂	☂	☀	☀	☀	☂	☂	☀	☂	☂	☂	☀	☀	☂	☀	☂	☂	☀	☂	☂	☂	☂	☀	☂	☂	
	夜	☂	☂	★	★	★	★	★	☂	★	☂	☂	☂	☂	★	★	★	★	★	☂	★	★	☂	★	☂	☂	☂	☂	☂	★	
	57年	昼	☂	☂	☀	☂	☀	☀	☀	☂	☀	☀	☀	☂	☀	☂	☀	☀	☀	☀	☀	☀	☂	☂	☂	☂	☂	☂	☀	☀	
	夜	☂	★	☂	☂	★	☂	★	☂	☂	★	★	☂	☂	☂	★	★	★	★	★	☂	★	★	☂	☂	☂	☂	☂	★	★	
月	58年	昼	☂	☂	☂	☂	☂	☂	☂	☂	☂	☂	☀	☂	☂	☂	☂	☂	☂	☂	☂	☂	☂	☂	☀	☀	☂	☂	☂	★	
	夜	☂	☂	☂	★	★	★	★	☂	★	★	★	★	☂	★	★	☂	☂	☂	☂	☂	★	☂	★	★	★	☂	☂	☂	★	
	累年平均	昼	☂	☂	☀	☀	☂	☀	☀	☂	☀	☀	☀	☀	☀	☀	☀	☀	☀	☀	☀	☀	☀	☂	☂	☂	☂	☂	☂	☀	
夜		★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★	★
气温旬間			2 -5	2 -5	2 -5	2 -5	2 -5	2 -5	2 -5	2 -5	2 -5	2 -5	2 -5	2 -5	3 -5	3 -5	3 -5	3 -5	3 -5	3 -5	3 -5	3 -5	3 -5	3 -5	3 -5	3 -5	4 -4	4 -4	4 -4	4 -4	4 -4
			☀ 43% ☂ 14% ☂ 4% ☂ 39% 2 -5										☀ 55% ☂ 10% ☂ 6% ☂ 29% 3 -5										☀ 40% ☂ 16% ☂ 4% ☂ 40%								

☂ sunny day

☂ night

☂ cloudy

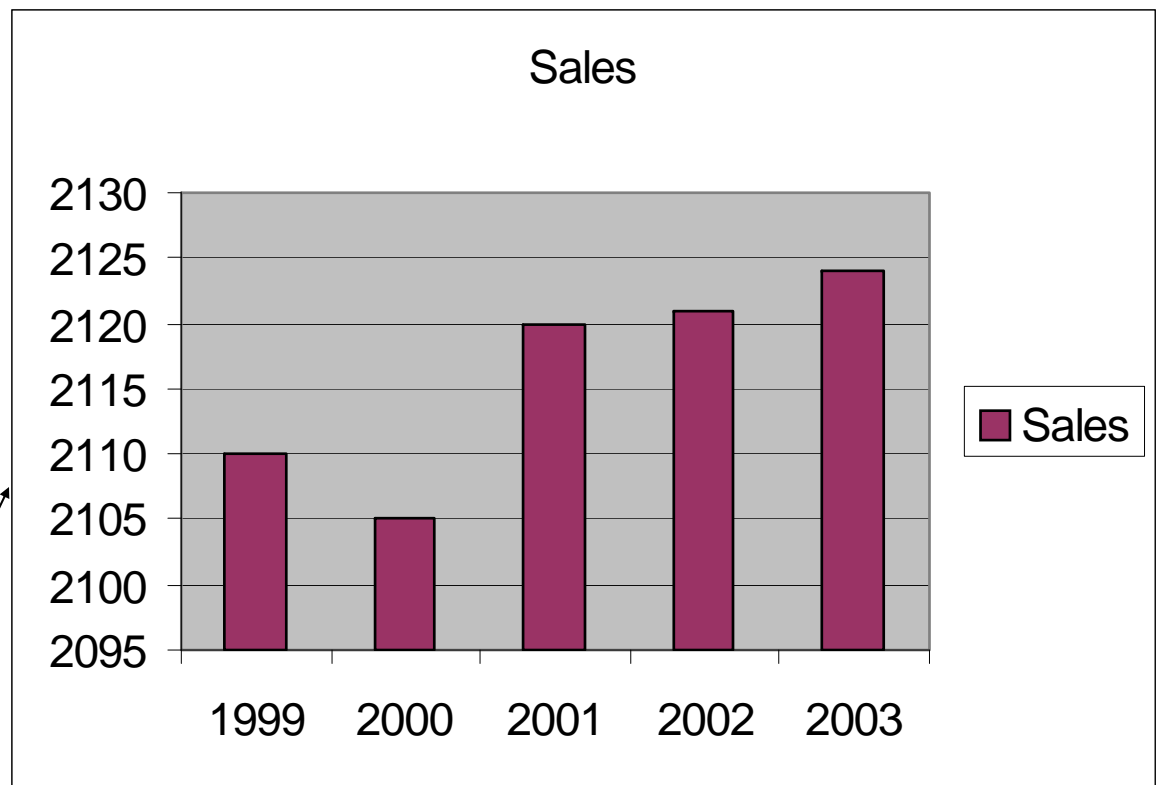
☂ rain

☂ snow

# Bad Visualization:

## Spreadsheet with misleading Y -axis

Year	Sales
1999	2110
2000	2105
2001	2120
2002	2121
2003	2124

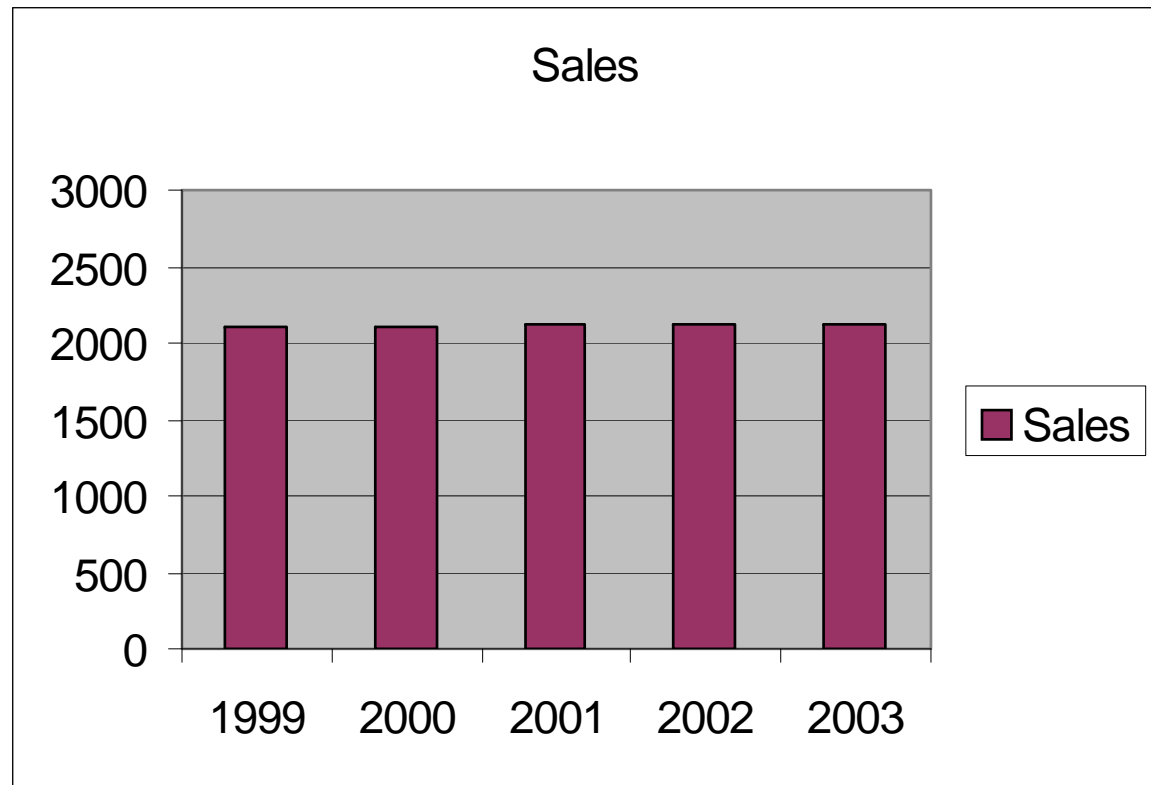


Y-Axis scale gives impression of big change



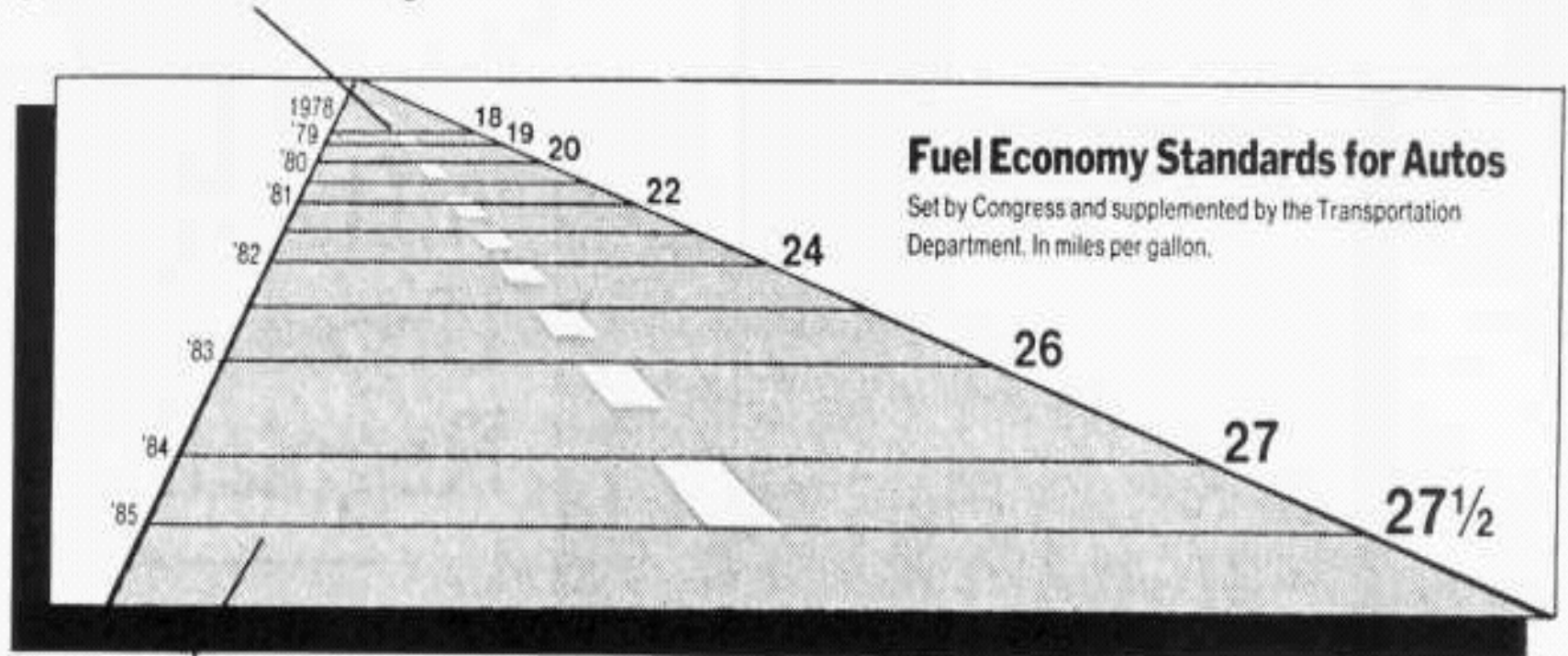
# Better Visualization

Year	Sales
1999	2110
2000	2105
2001	2120
2002	2121
2003	2124



Axis from 0 to 2000 scale gives  
correct impression of small change

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

Lie Factor=14.8

*New York Times*, August 9, 1978, p. D-2.

(E.R. Tufte, "The Visual Display of Quantitative Information", 2nd edition)

# Lie Factor

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

$$= \frac{\frac{(5.3 - 0.6)}{0.6}}{\frac{(27.5 - 18.0)}{18}} = \frac{7.833}{0.528} = 14.8$$

Tufte requirement:  $0.95 < \text{Lie Factor} < 1.05$

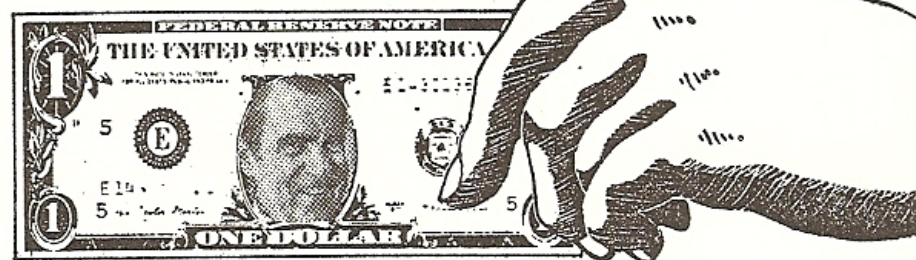
(E.R. Tufte, "The Visual Display of Quantitative Information", 2nd edition)



**1963 — KENNEDY: 94c**



**1968 — JOHNSON: 83c**



**1973 — NIXON: 64c**

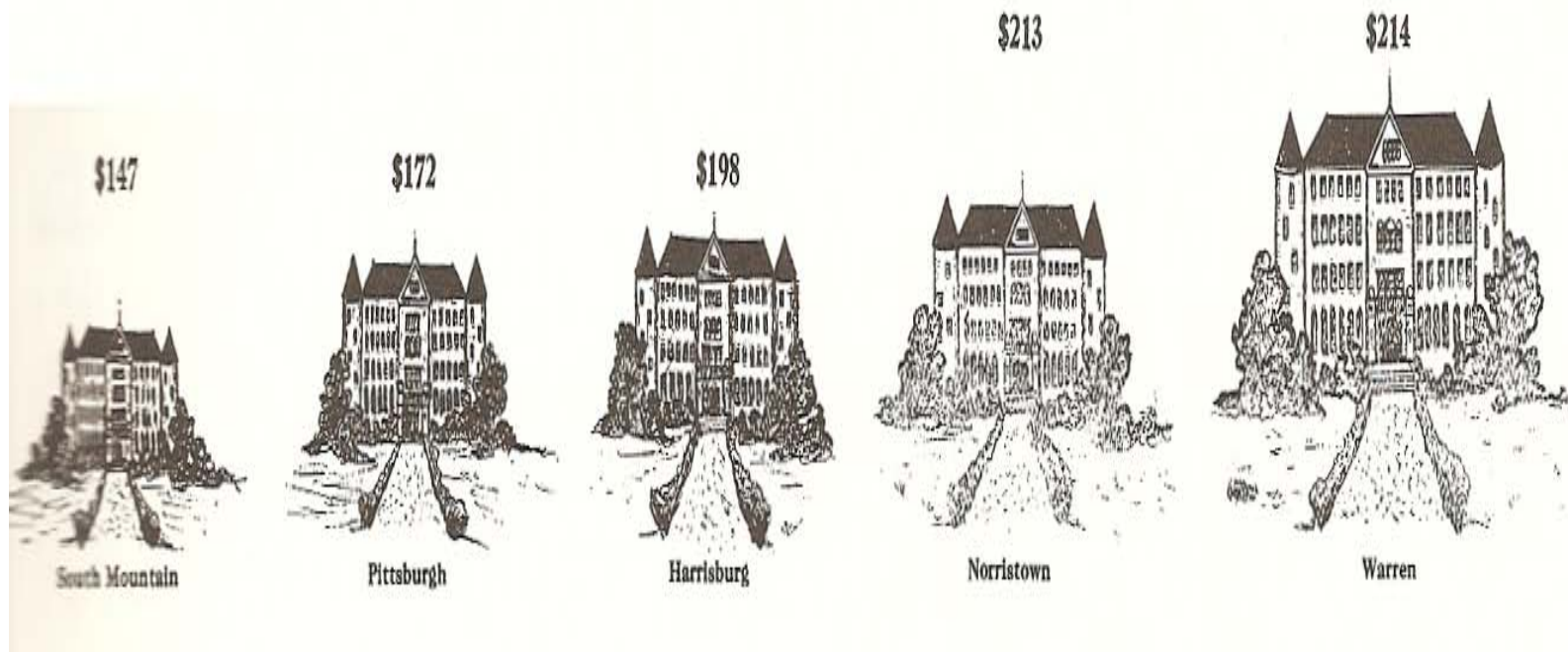


**1978 — CARTER: 44c**  
(August)

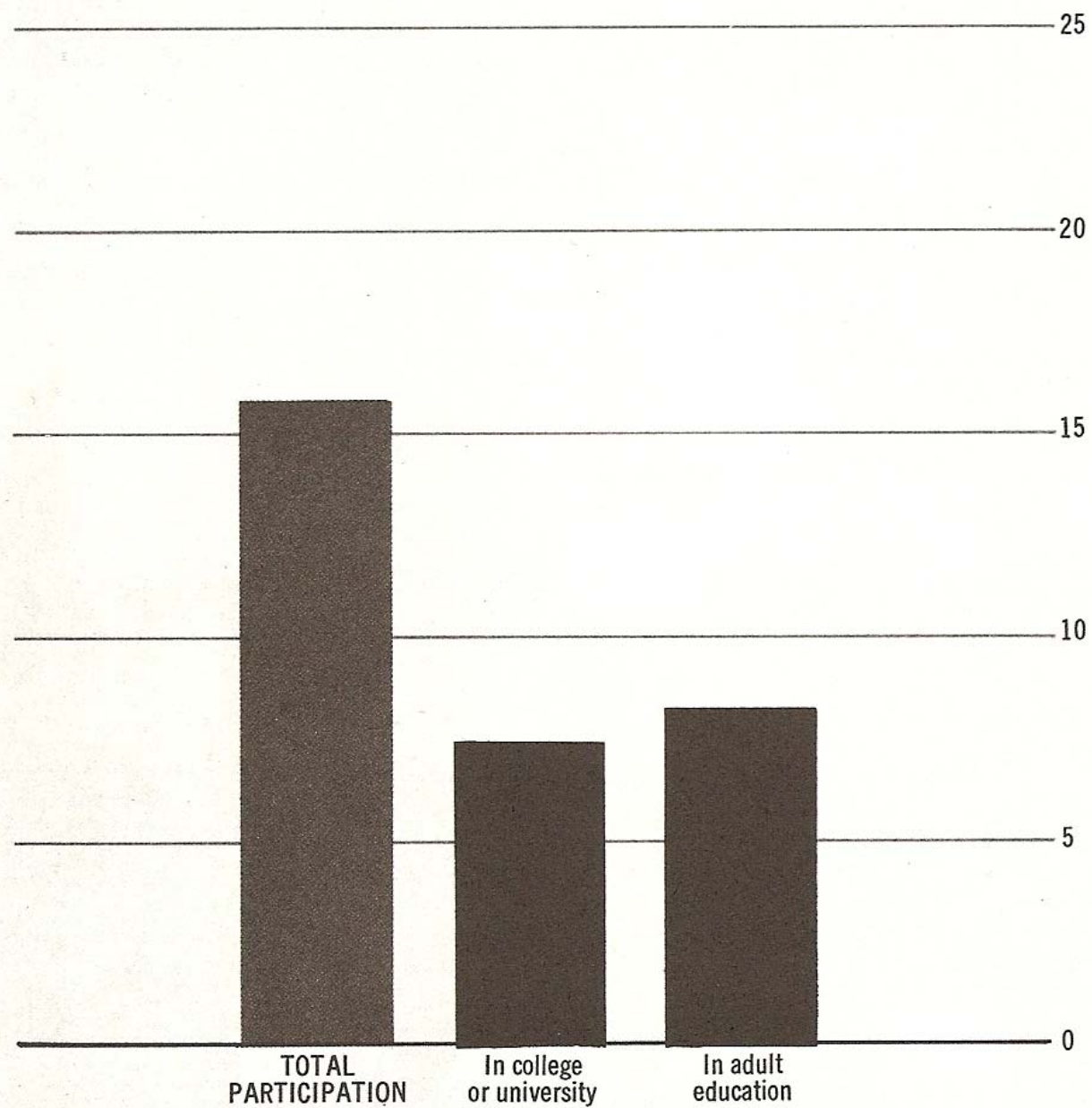
# Purchasing Power of the Diminishing Dollar

Source: Labor Department



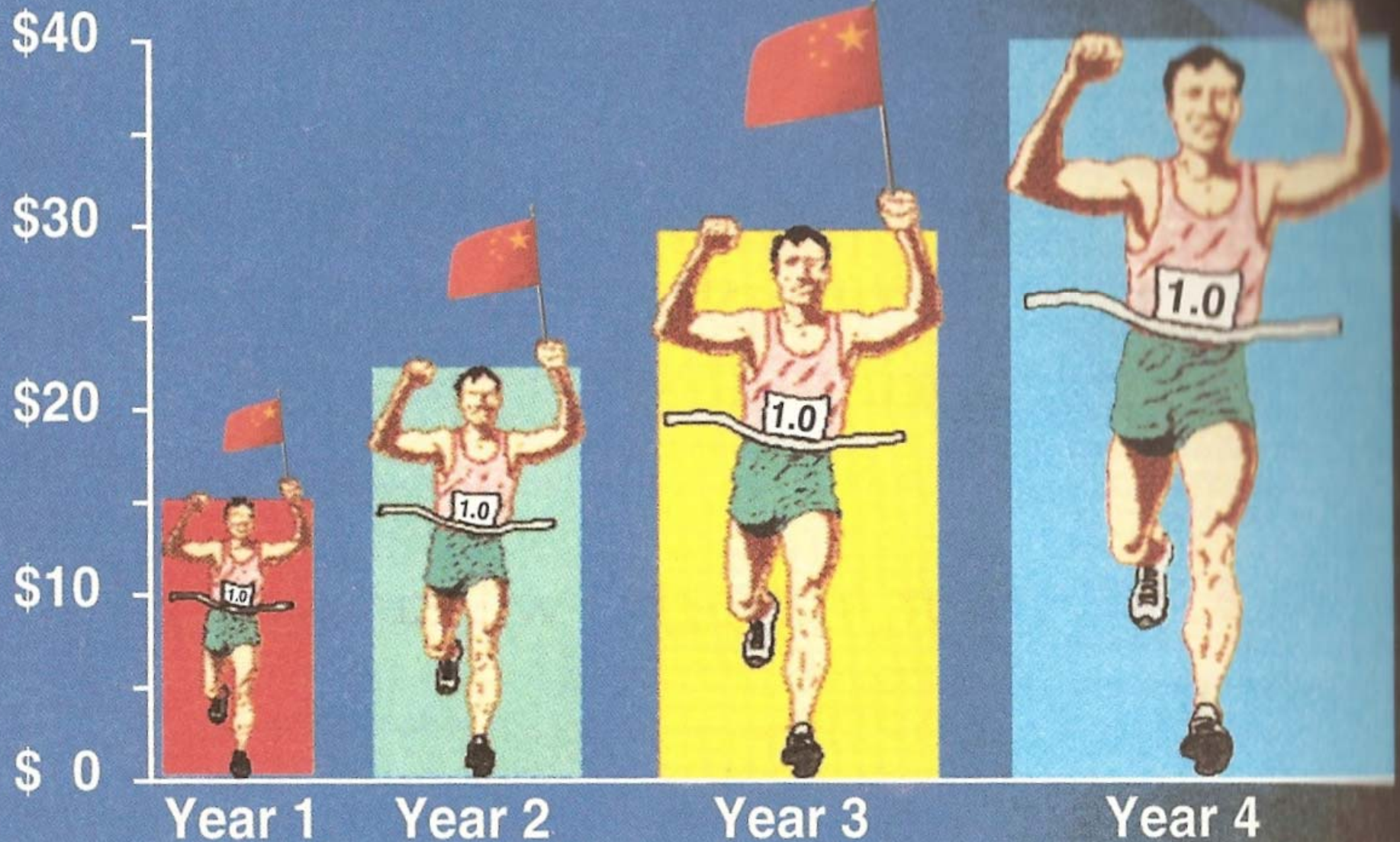


(E. Tufte, "The Visual Display of Quantitative Information", pg. 55) <sup>37</sup>



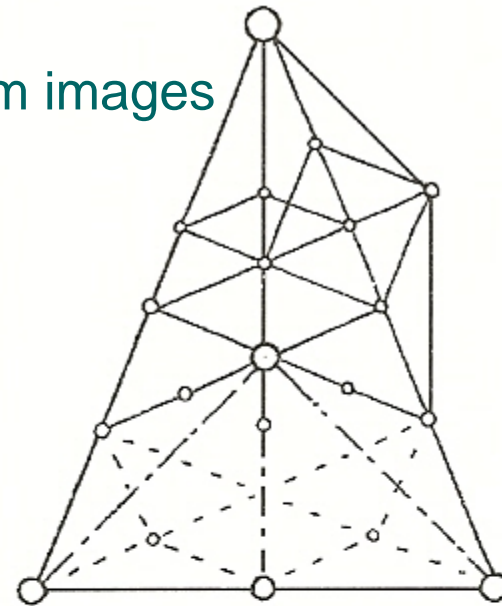
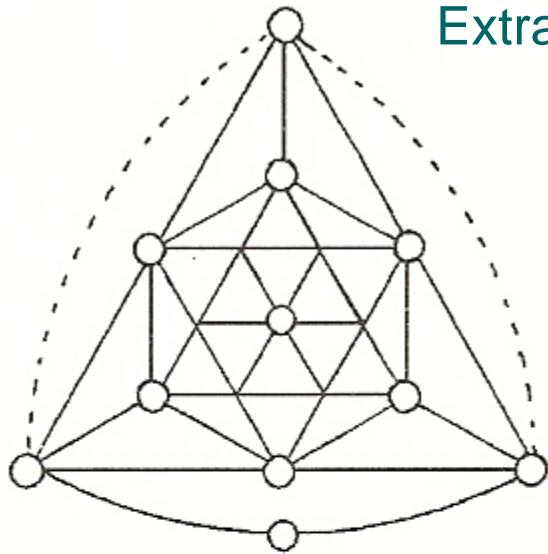


# Revenue Growth Forecasts



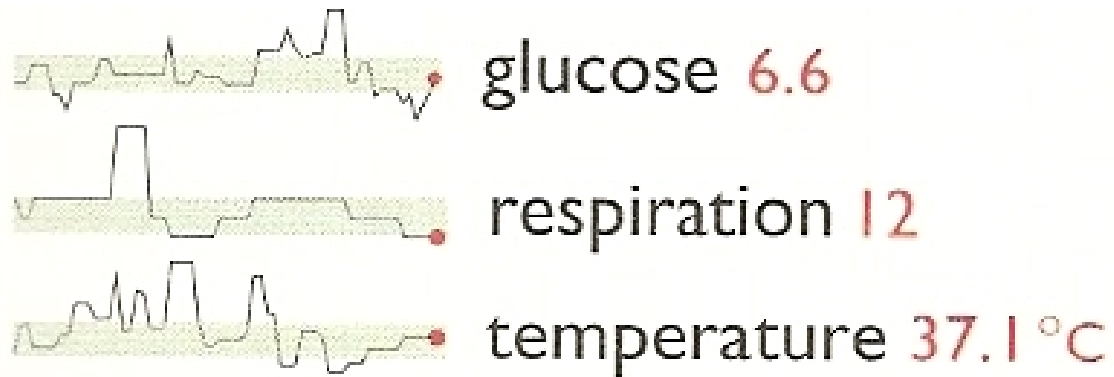


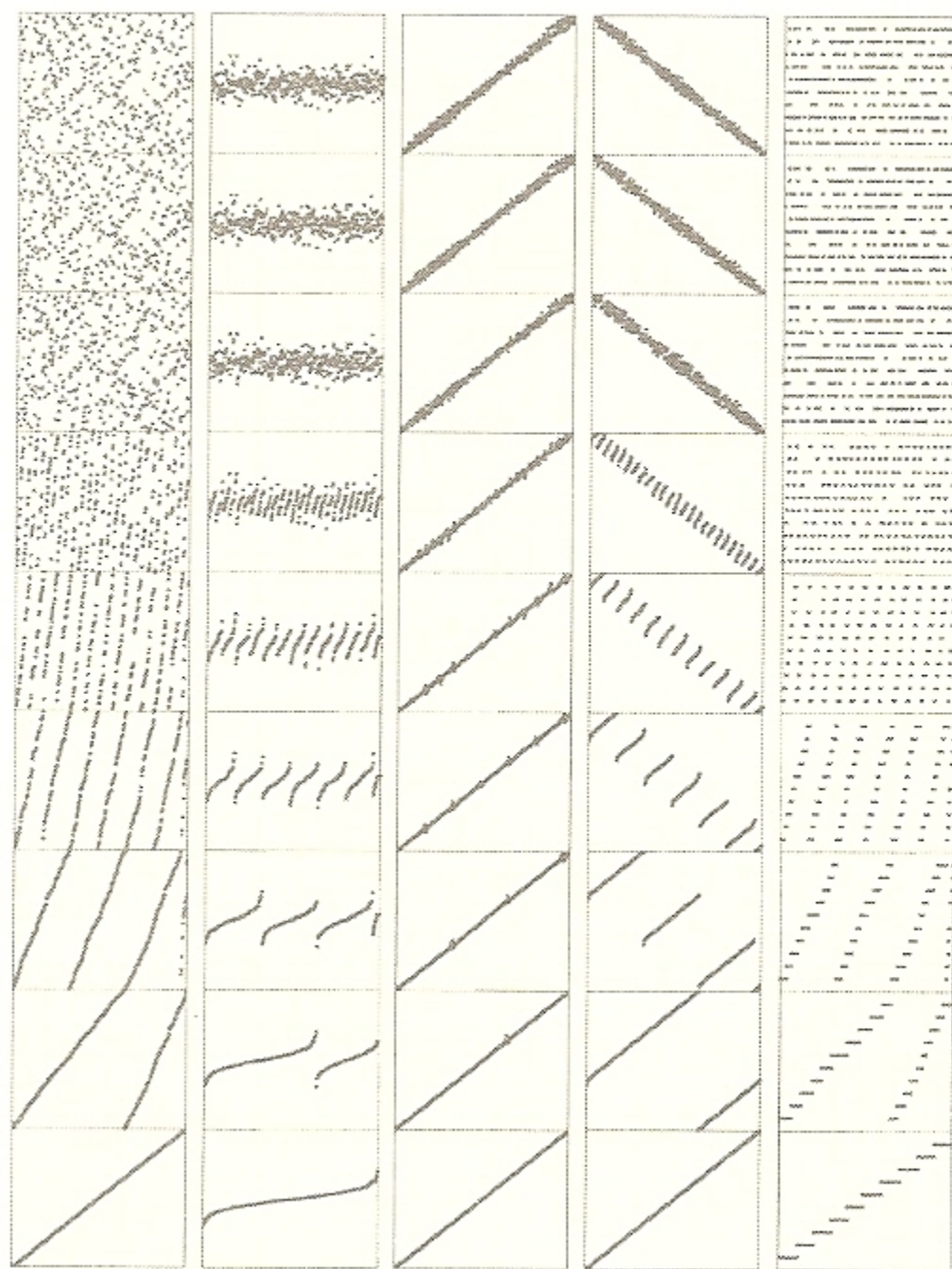
## Extracting graphs from images



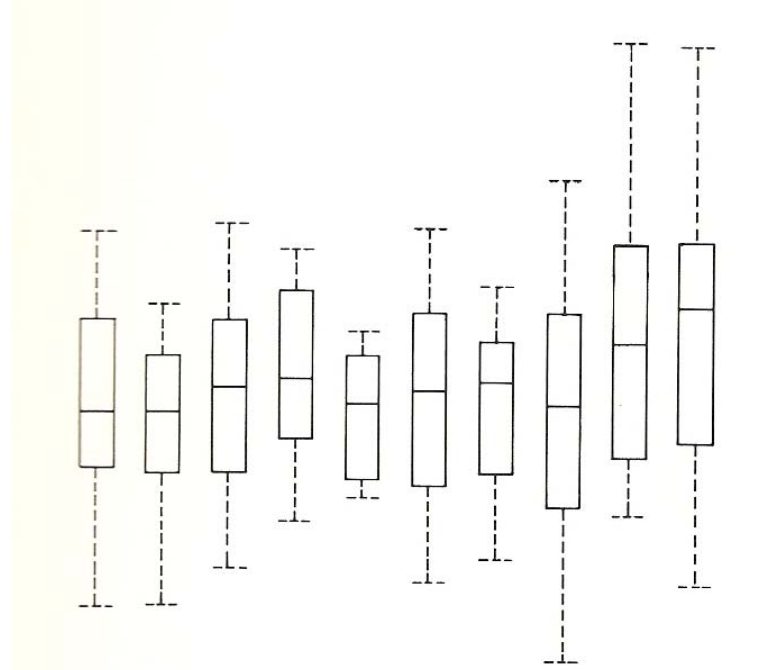
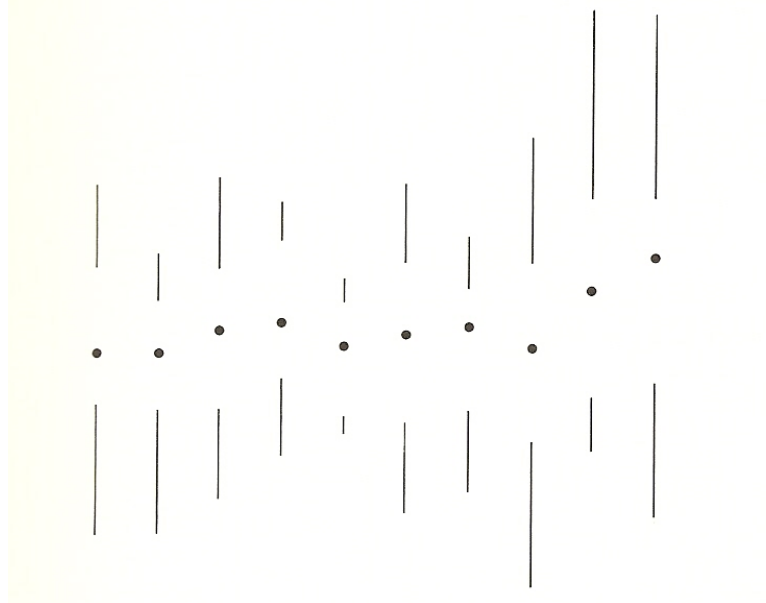


# Sparklines: Word-Sized Graphics

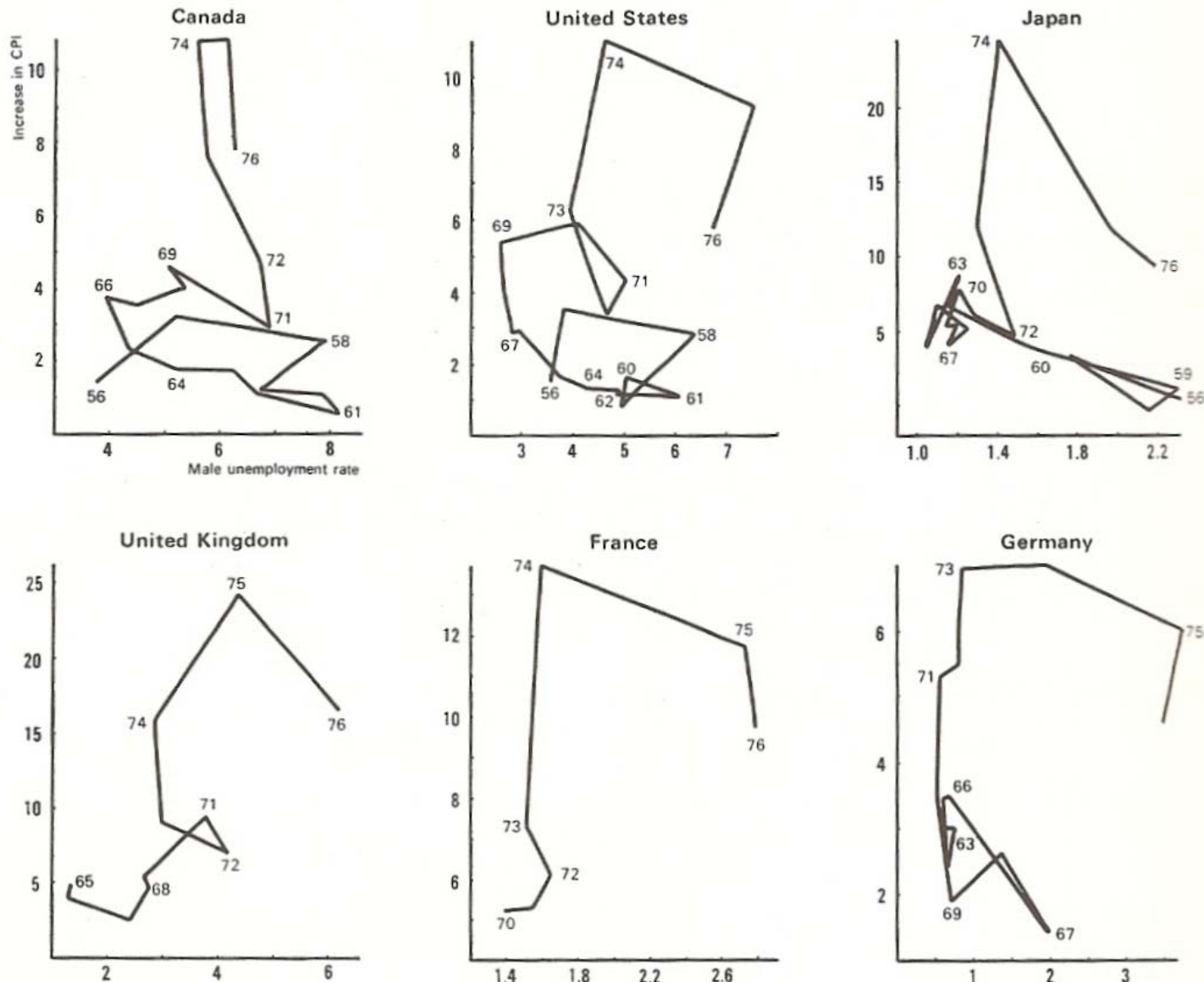




## Re-designing box plots



## Phillips curve plots



(E. Tufte, "The Visual Display of Quantitative Information", pg. 48)

# Bivariate rugplot



# Tufte's Principles of Graphical Excellence

Give the viewer

- the greatest number of ideas
- in the shortest time
- with the least ink in the smallest space
  
- maximize the proportion of a graphic's ink devoted to the non-redundant display of data
- tell the truth about the data