Visualization
and
Data Mining

Outline

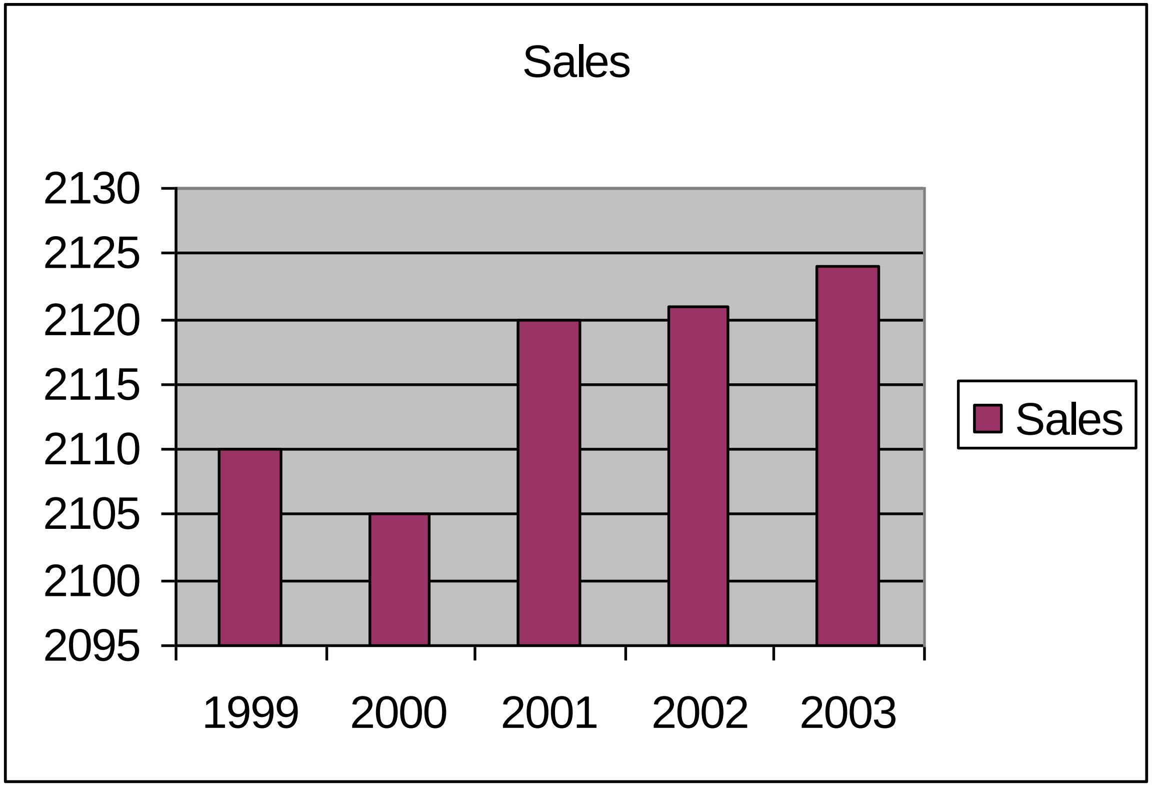
- Graphical excellence and lie factor
- Representing data in 1-D, 2-D, and 3-D
- Representing data in 4+ dimensions
  - Parallel coordinates
  - Scatterplots
  - Stick figures

Visualization Role

- Support interactive exploration
- Help in result presentation
- Disadvantage: requires human eyes
- Can be misleading

Bad Visualization: Spreadsheet

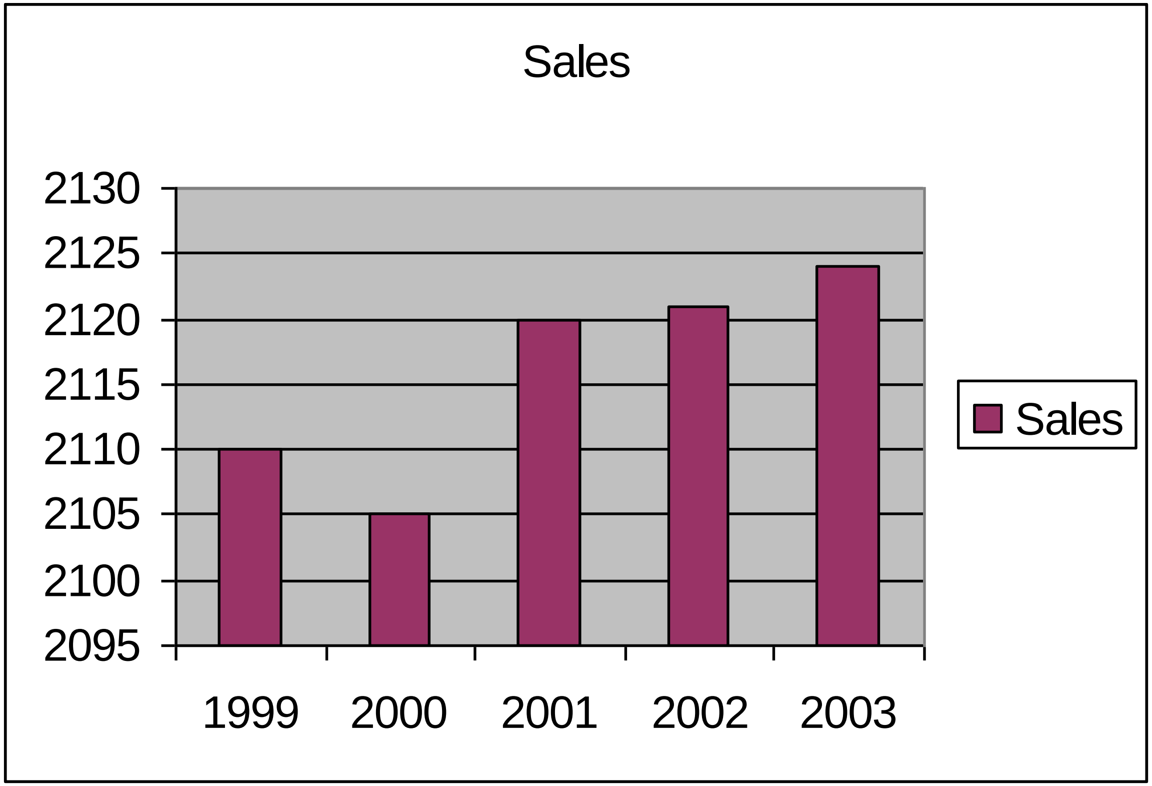
Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



What is wrong with this graph?

Bad Visualization:
Spreadsheet with misleading Y –axis

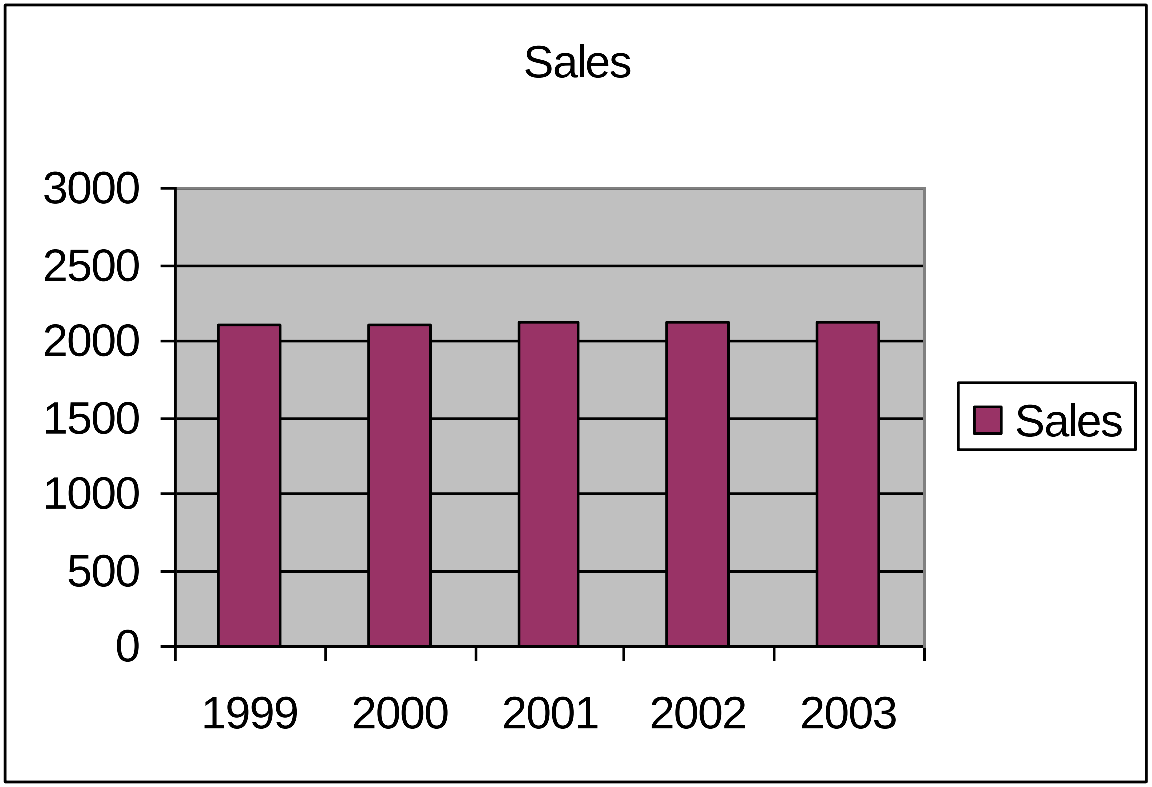
Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



Y-Axis scale gives **WRONG** impression of big change

Better Visualization

Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



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

Lie Factor

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

$$= \frac{\dfrac{(20-10)}{10}}{\dfrac{(2120-2105)}{2105}} = \frac{0.5}{0.007125} = 70.18$$

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

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

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.

- Tell the truth about the data!

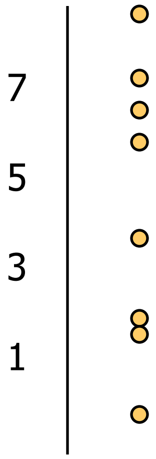
(E.R. Tufte, "The Visual Display of Quantitative Information", 2nd edition)

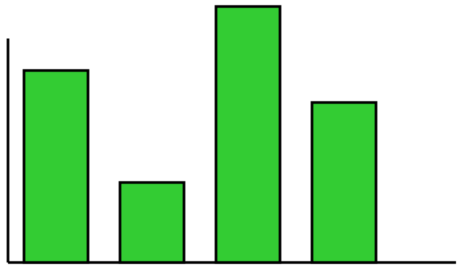
Visualization Methods

- Visualizing in 1-D, 2-D and 3-D
  - well-known visualization methods
- Visualizing more dimensions
  - Parallel Coordinates
  - Other ideas

1-D (Univariate) Data

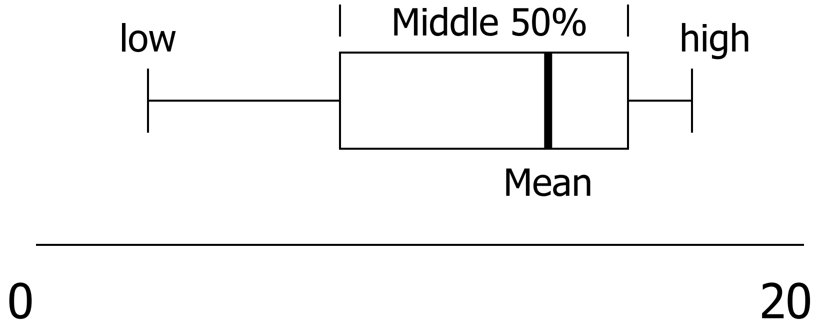
- Representations





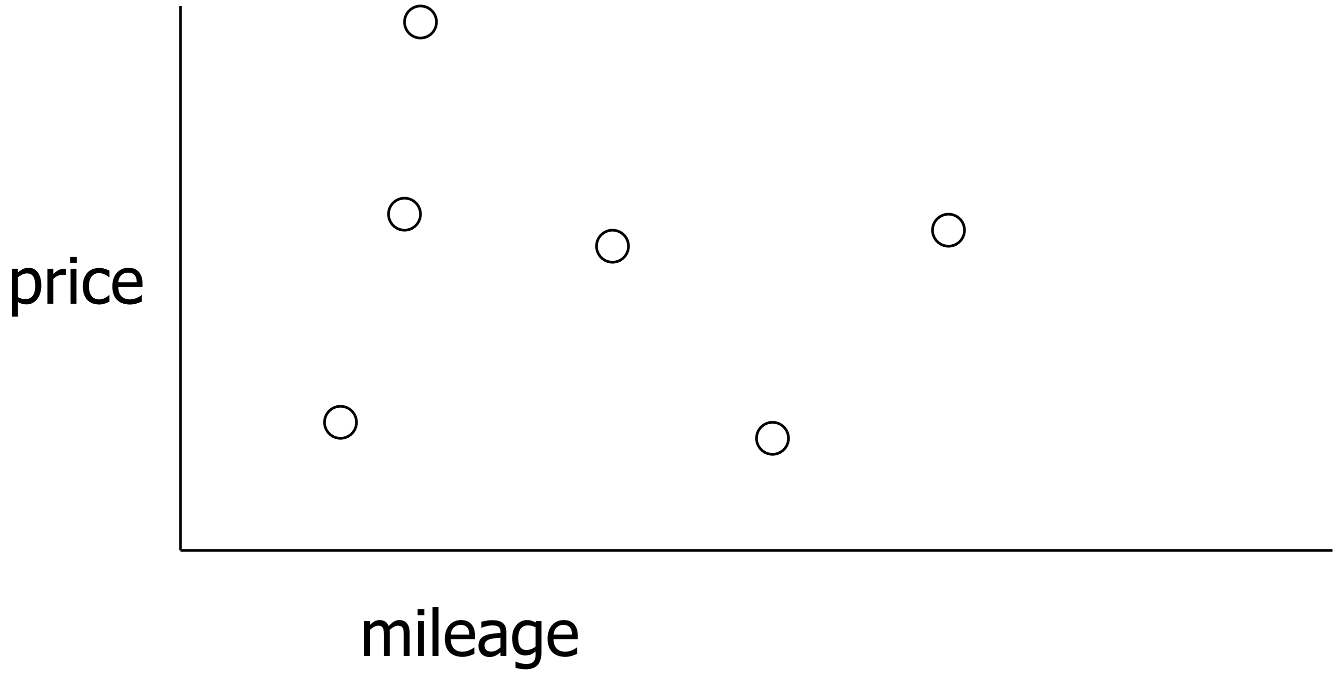
Histogram

Tukey box plot

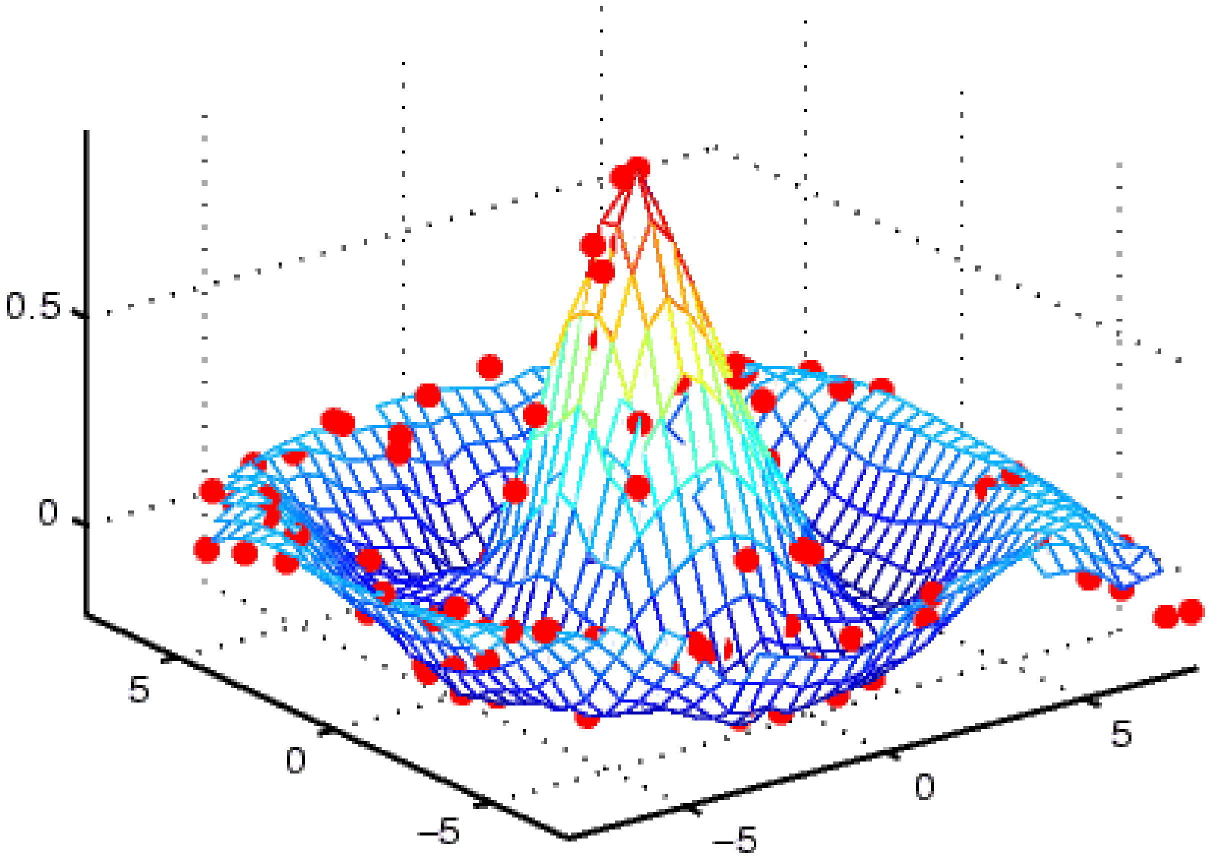


2-D (Bivariate) Data

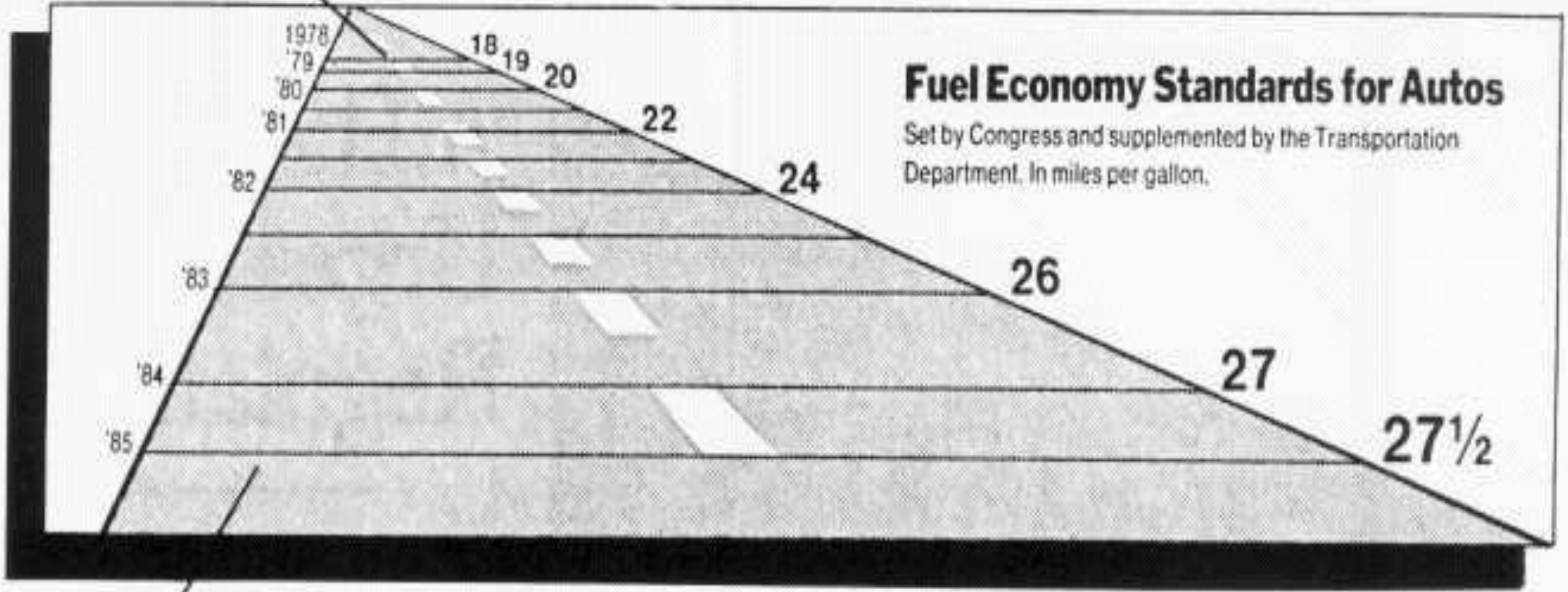
- Scatter plot, …



3-D Data (projection)



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



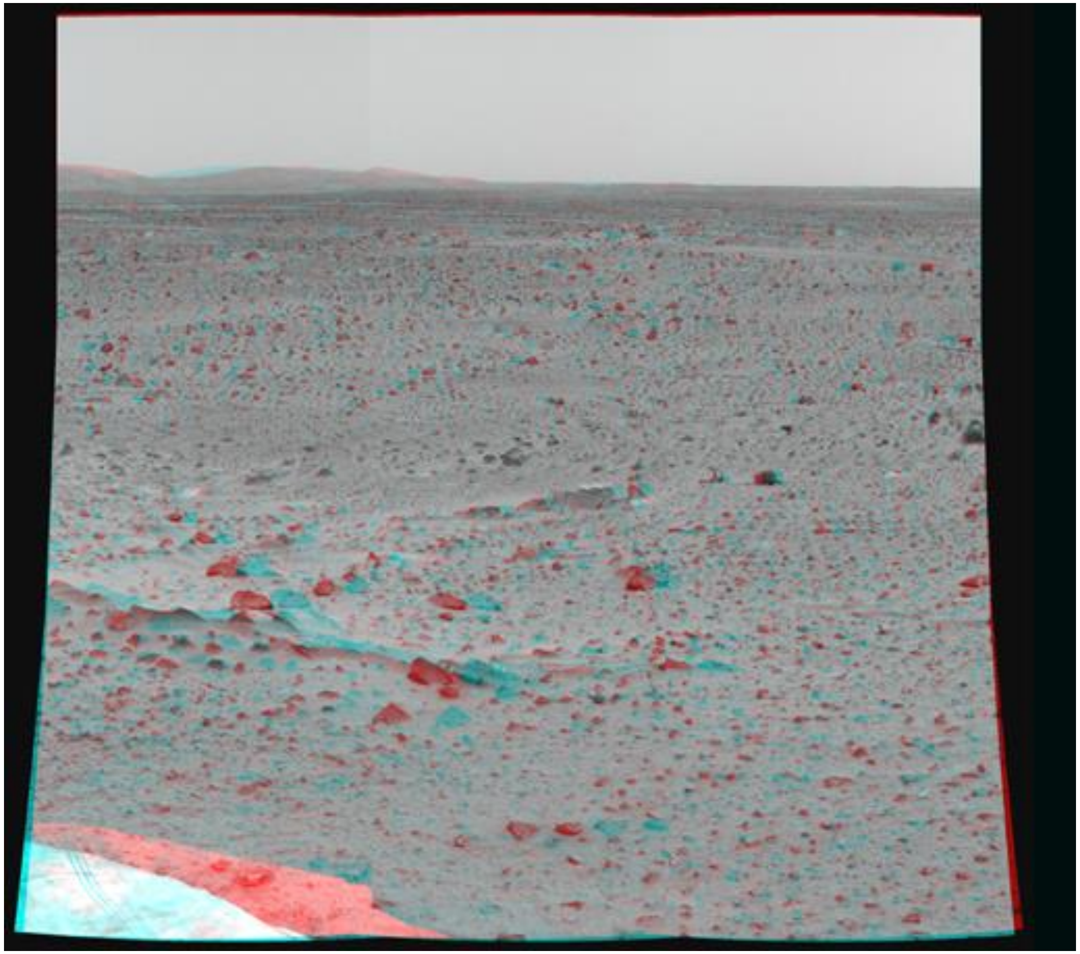
This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

Lie Factor=14.8

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

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

3-D image
(requires 3-D blue and red glasses)



Taken by Mars Rover Spirit, Jan 2004

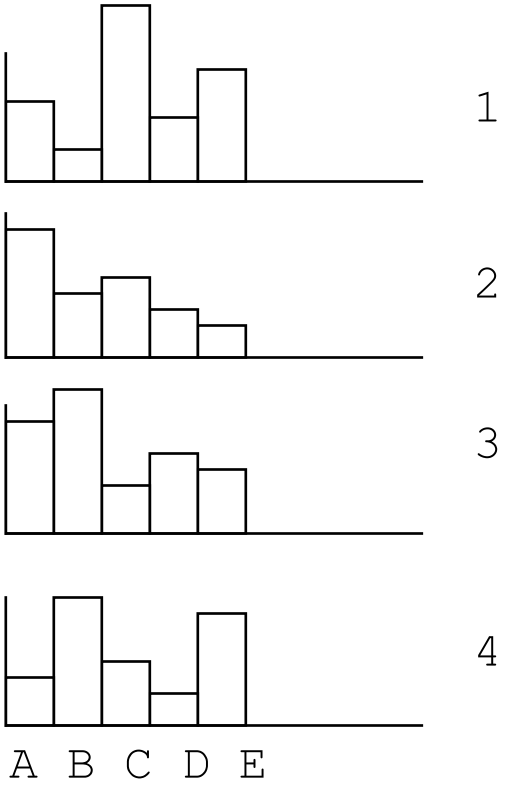
Visualizing in 4+ Dimensions

- Scatterplots
- Parallel Coordinates
- Chernoff faces
- 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



Problem: does not show correlations

Scatterplot Matrix

Represent each possible pair of variables in their own 2-D scatterplot (car data)

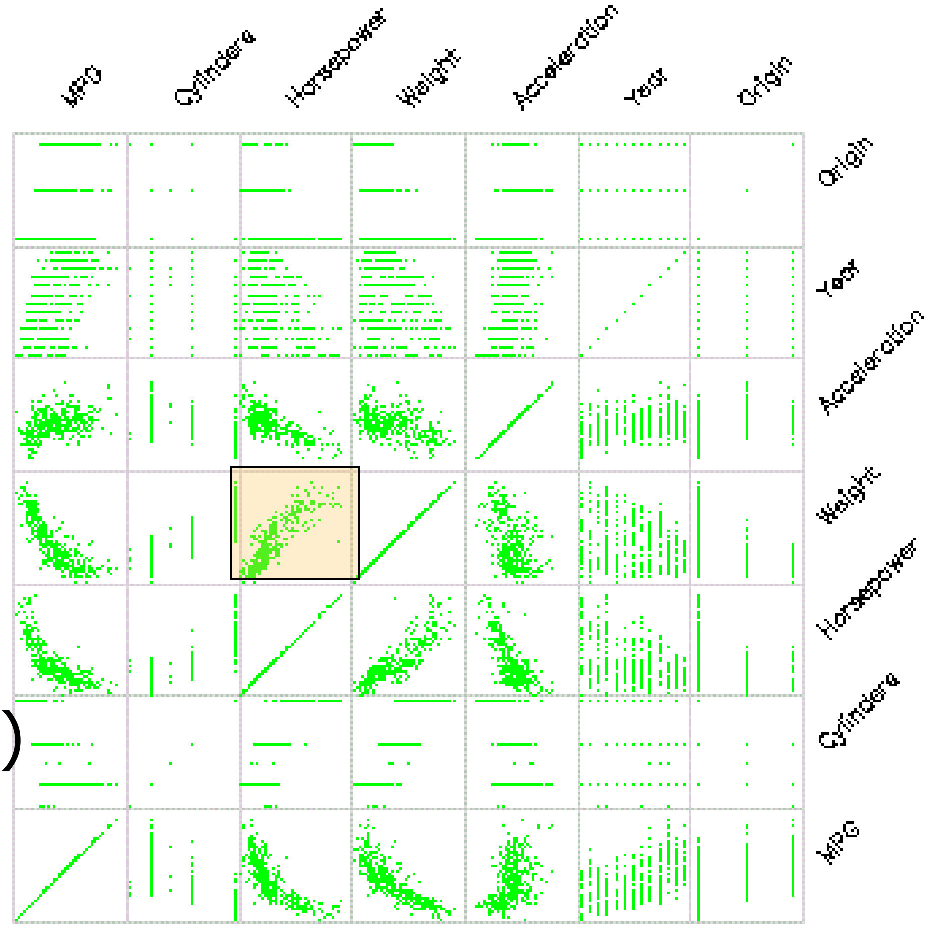
Q: Useful for what?

A: linear correlations

(e.g. horsepower & weight)

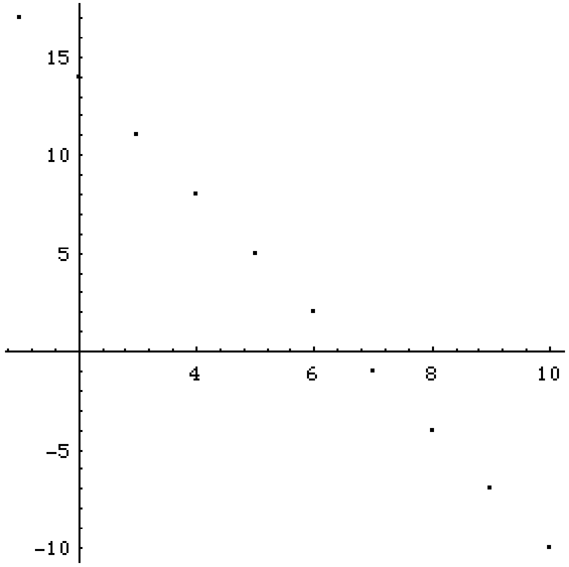
Q: Misses what?

A: multivariate effects

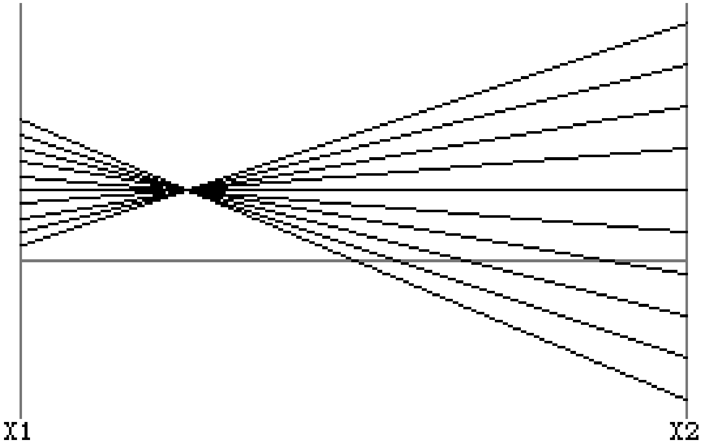


Parallel Coordinates

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



Dataset in a Cartesian coordinates



Same dataset in parallel coordinates

Invented by
Alfred Inselberg
while at IBM, 1985

Example: Visualizing Iris Data

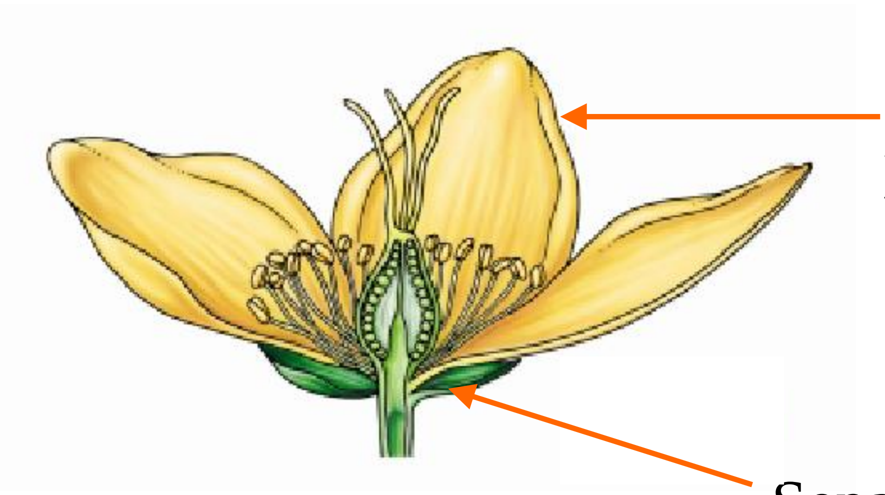
Iris setosa

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

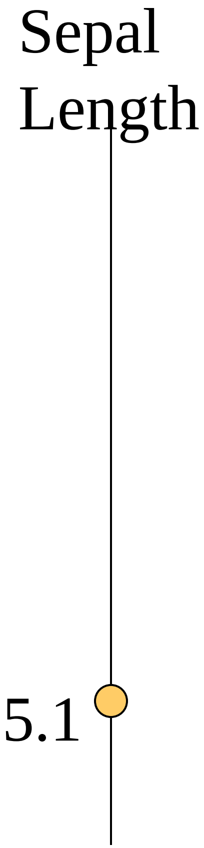
Flower Parts



Petal, a non-reproductive part of the flower

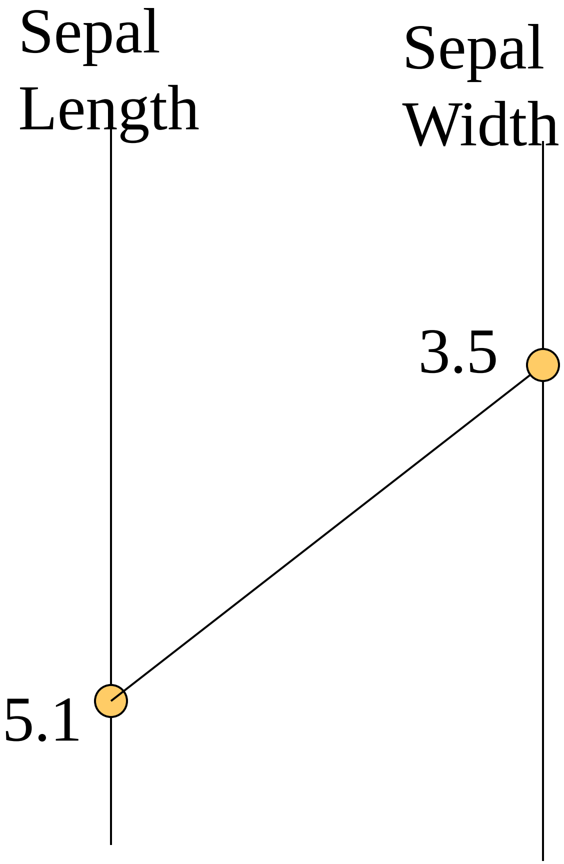
Sepal, a non-reproductive part of the flower

Parallel Coordinates



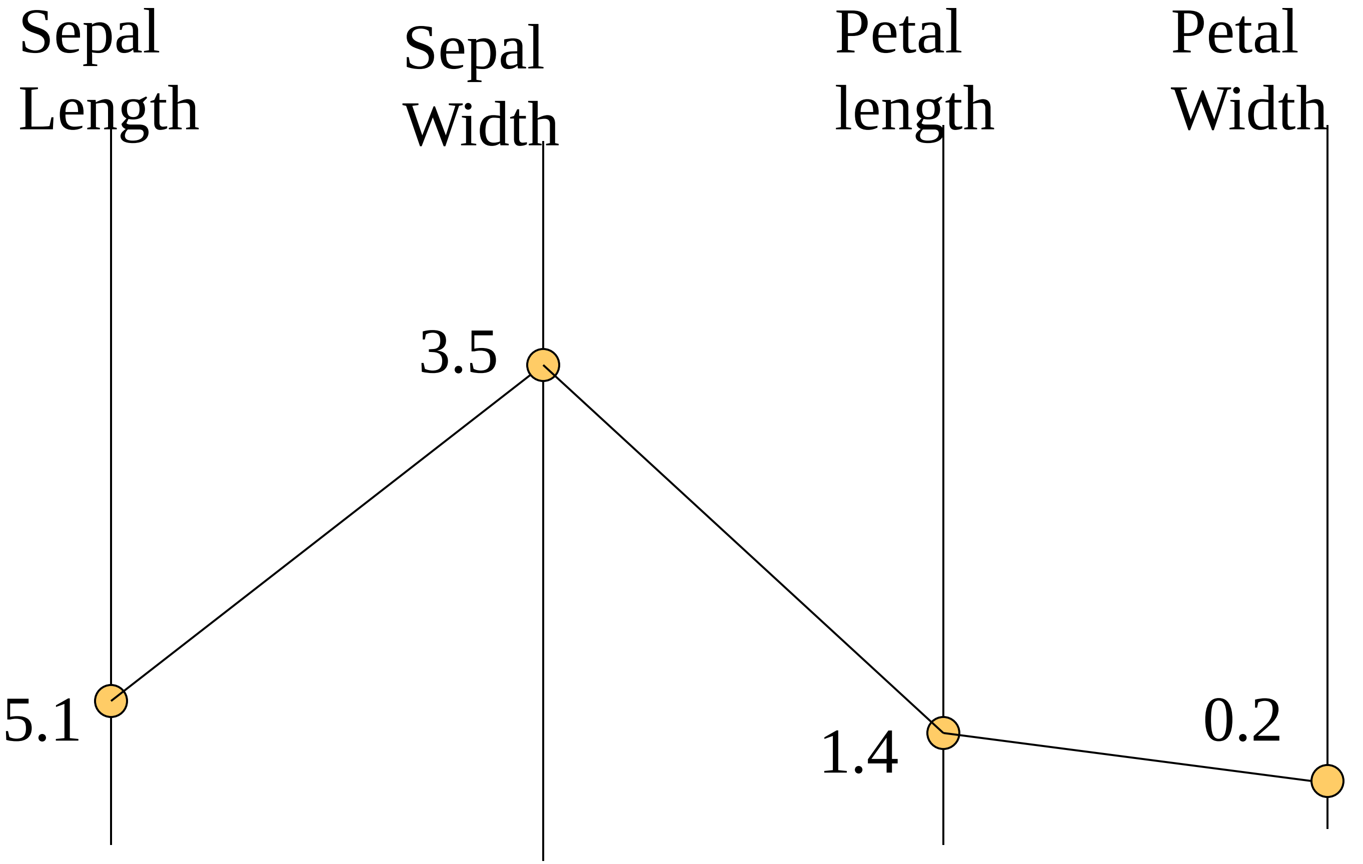
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Parallel Coordinates: 2 D



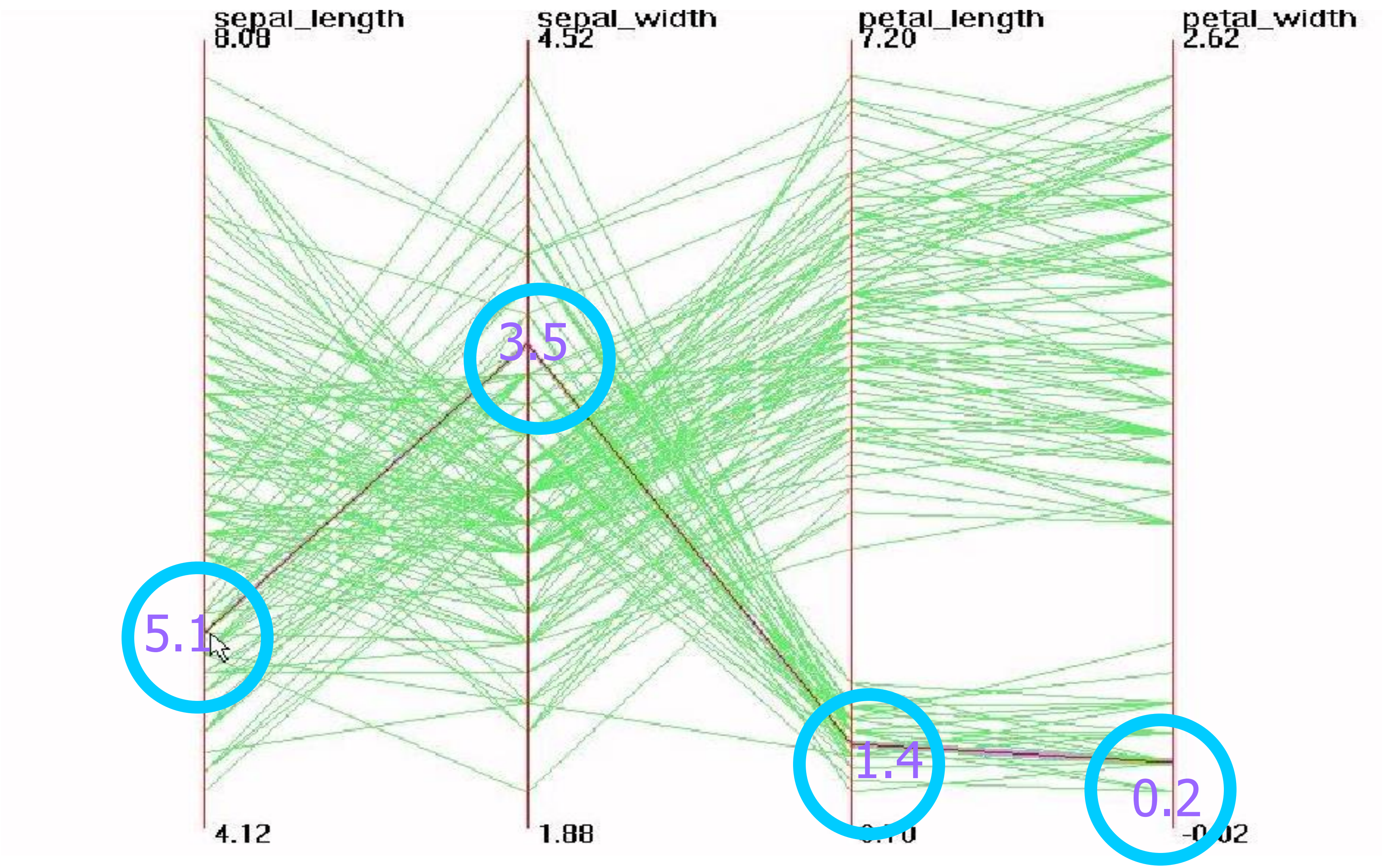
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Parallel Coordinates: 4 D



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

Parallel Visualization of Iris data



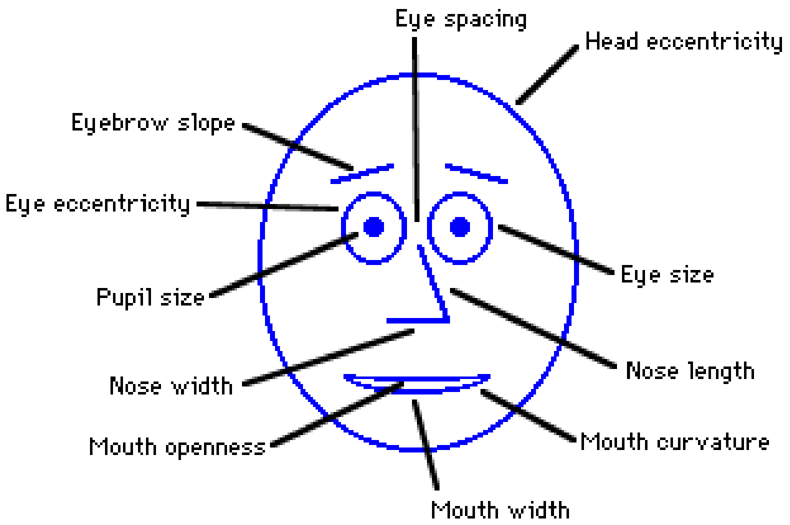
Parallel Visualization Summary

- Each data point is a line
- Similar points correspond to similar lines
- Lines crossing over correspond to negatively correlated attributes
- Interactive exploration and clustering

- Problems: order of axes, limit to ~20 dimensions

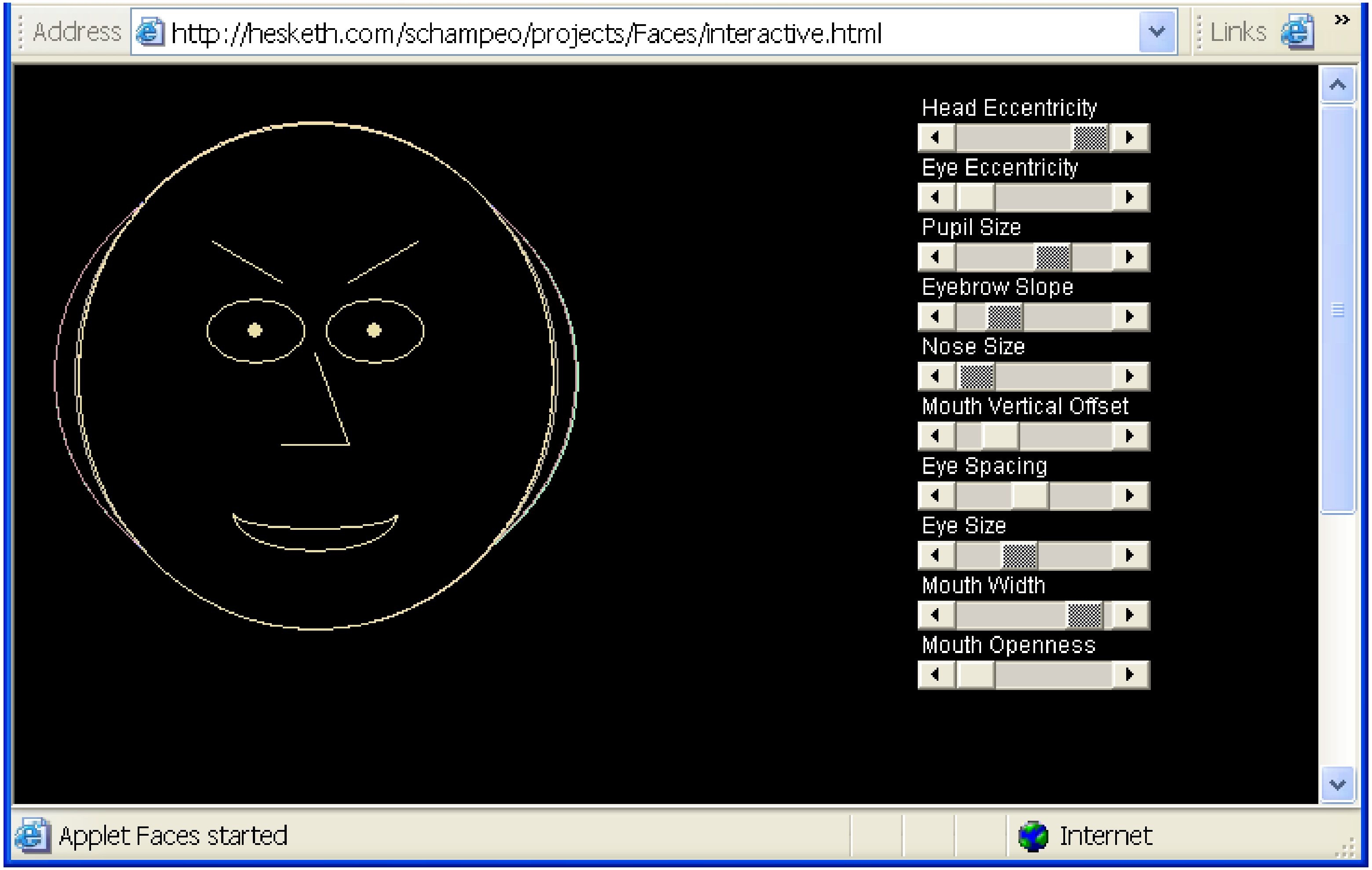
Chernoff Faces

Encode different variables' values in characteristics of human face



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

Interactive Face

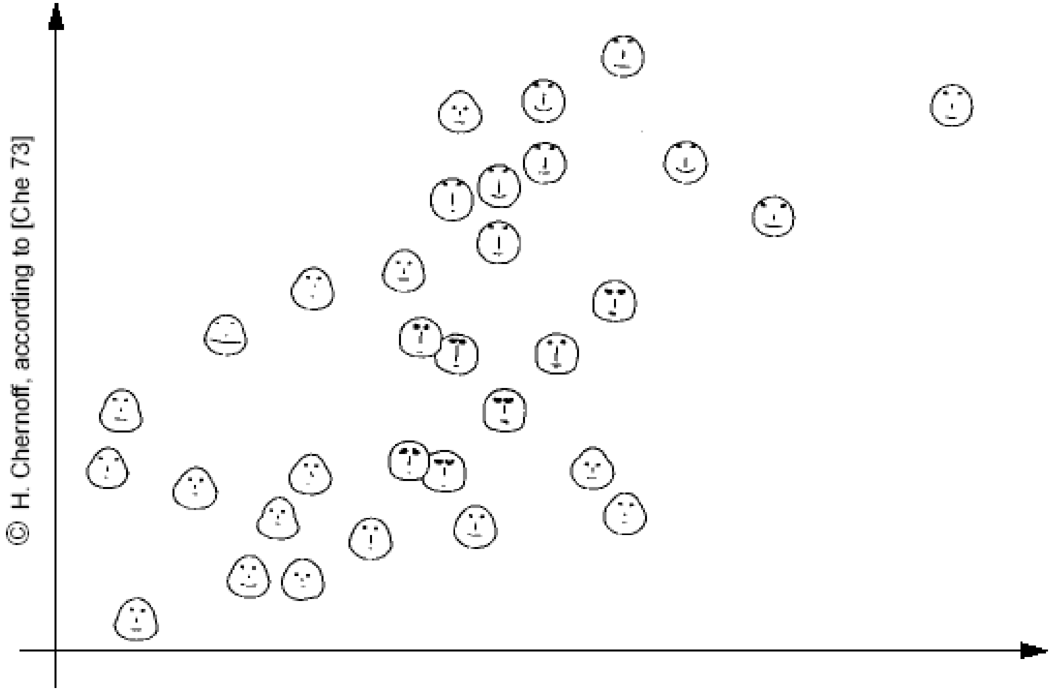


Address http://hesketh.com/schampeo/projects/Faces/interactive.html

Applet Faces started

Chernoff faces, example

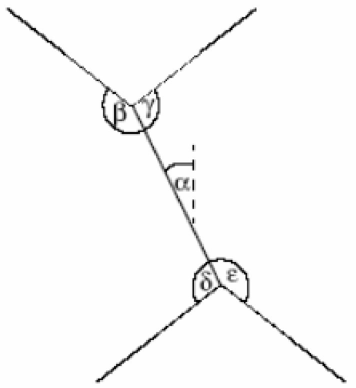
Chernoff-Faces [Che 73, Tuf 83]



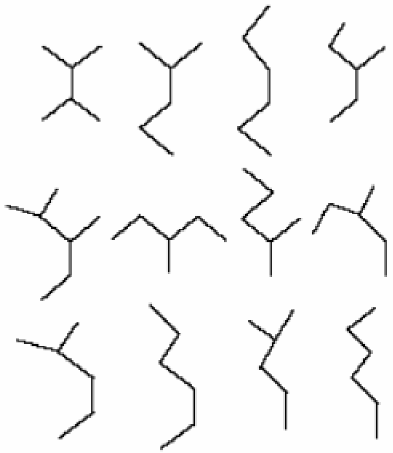
© H. Chernoff, according to [Che 73]

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

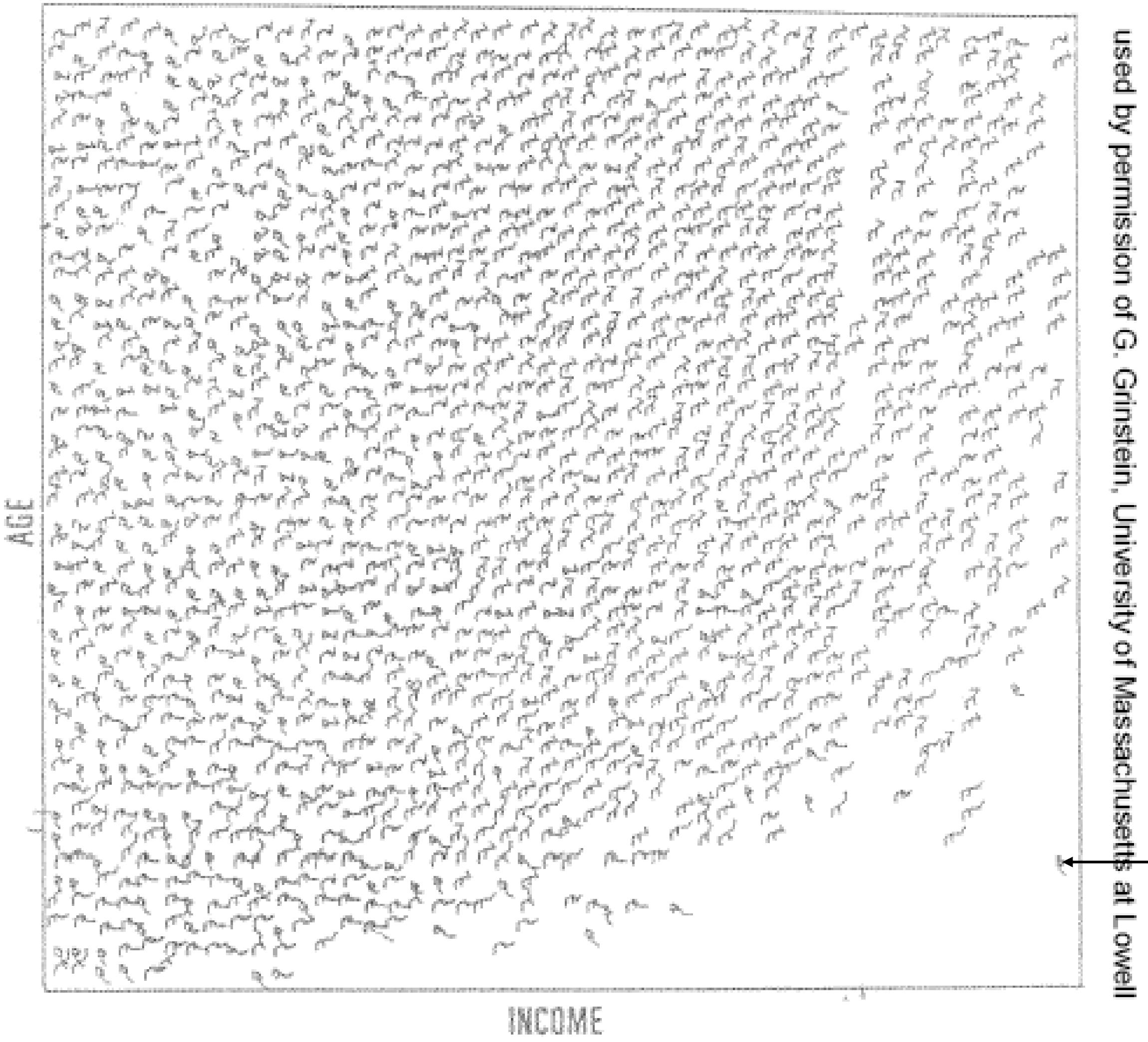


Stick Figure Icon



A Family of Stick Figures

Stick figures, example



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

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

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

Note also a young woman with high income

Visualization software

Free and Open-source

- R + ggplot2
- Python + vizualization libraries
- Orange

- Many more – see:
  www.KDnuggets.com/software/visualization.html

Visualization Summary

- Many methods
- Visualization is possible in more than 3-D
- Aim for graphical excellence
- Tell the truth about the data