Data Mining: Data Exploration

Tom Claassen

Radboud University Nijmegen

What is data exploration?

- A preliminary exploration of the data to better understand its characteristics
- Key motivations of data exploration include
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans' abilities to recognize patterns
- Related to the area of Exploratory Data Analysis (EDA)
  - Created by statistician John Tukey
  - Seminal book is Exploratory Data Analysis by Tukey
  - A nice online introduction can be found in Chapter 1 of the NIST Engineering Statistics Handbook http://www.itl.nist.gov/div898/handbook/index.htm

Old Faithful

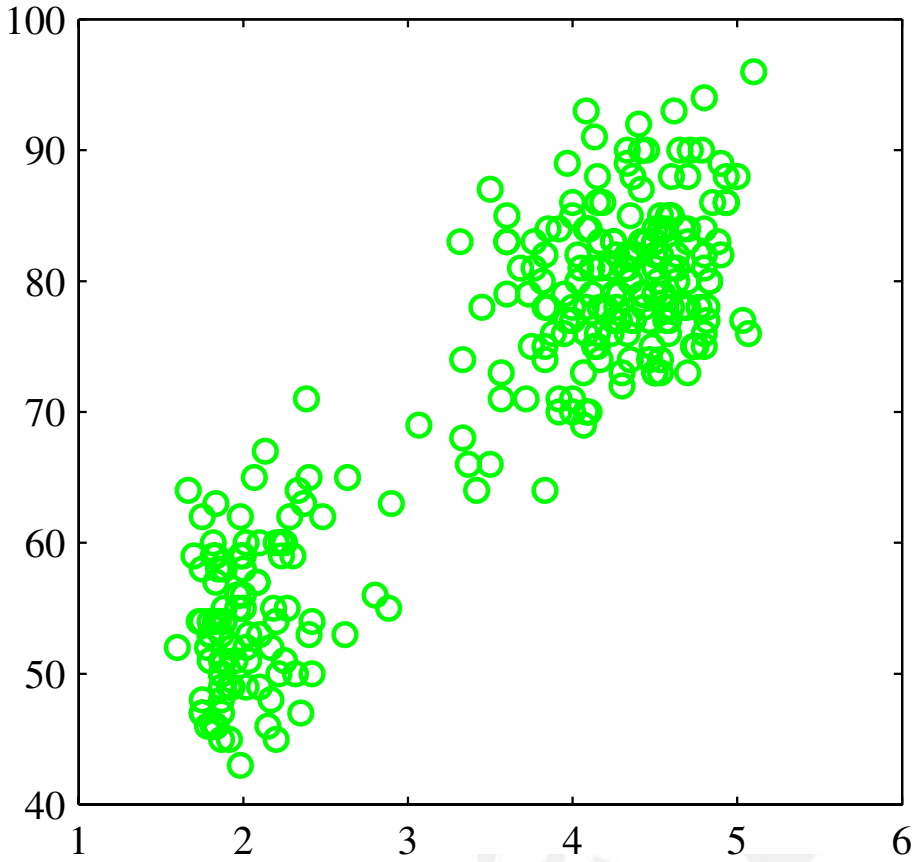
- famous geyser in Yellowstone National Park (Wyoming, US)
- very regular eruptions every 1-1.5 hrs.
- each eruption lasting 2-5min.

observation: after bigger eruptions it seems to take longer to the next eruption …

Data set analysis

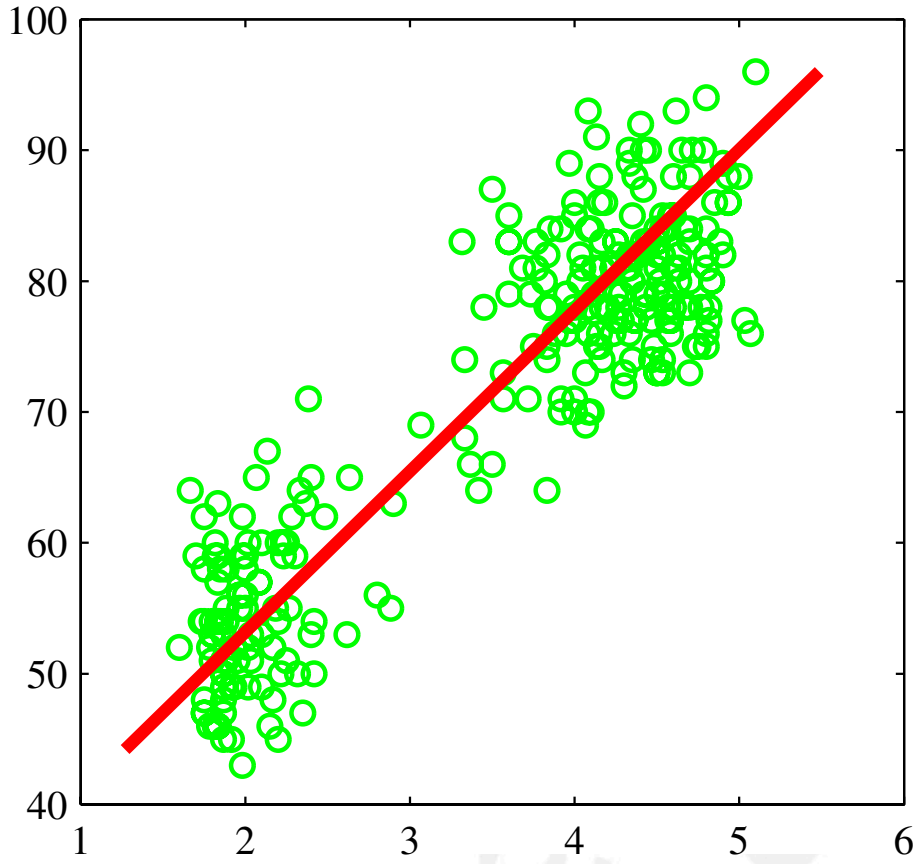
- possible explanations?
- can we understand what is really going on?
- how to tackle?

Old Faithful



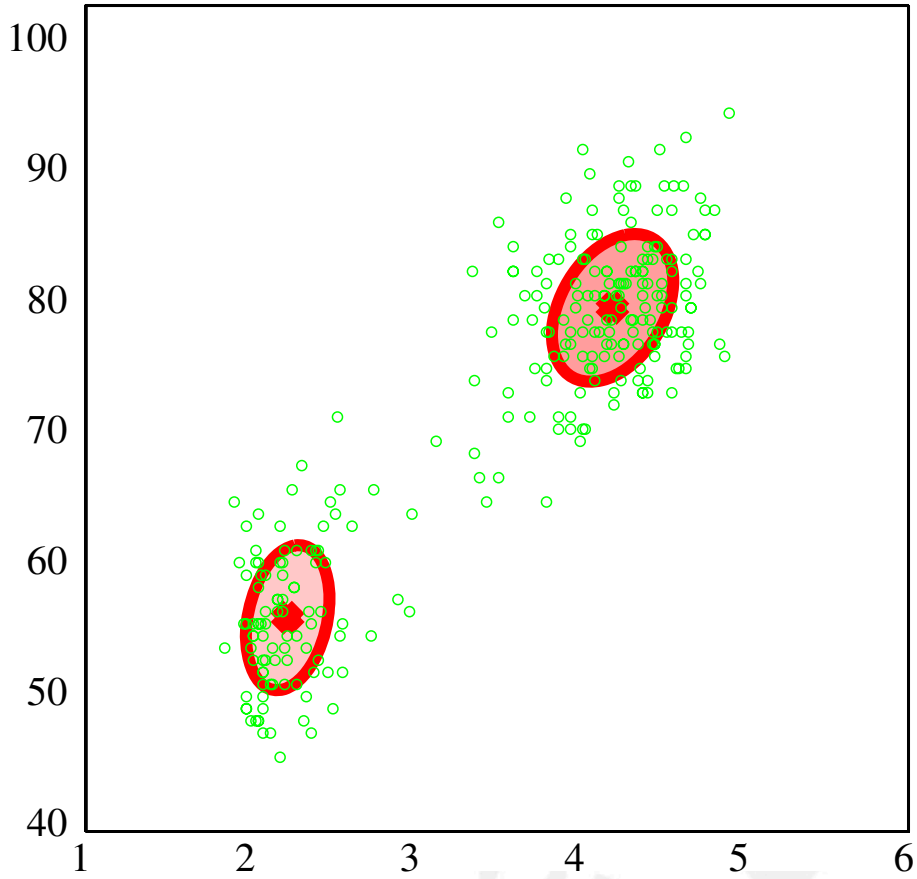
Duration of eruption in min. (horizontal axis) vs. time to next eruption (vertical)

Old Faithful



Duration of eruption in min. (horizontal axis) vs. time to next eruption (vertical)

Old Faithful

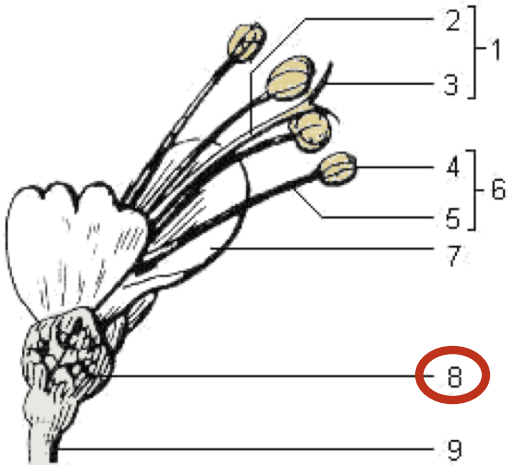


Duration of eruption in min. (horizontal axis) vs. time to next eruption (vertical)

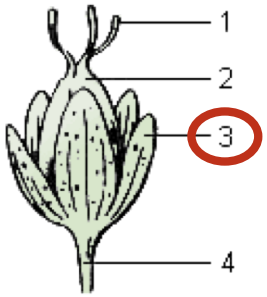
Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set
- Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
- From the statistician Douglas Fisher
- Three flower types (classes):
  - Setosa
  - Virginica
  - Versicolour
- Four (non-class) attributes
  - Sepal (kelkblad) width and length
  - Petal (bloemblad) width and length

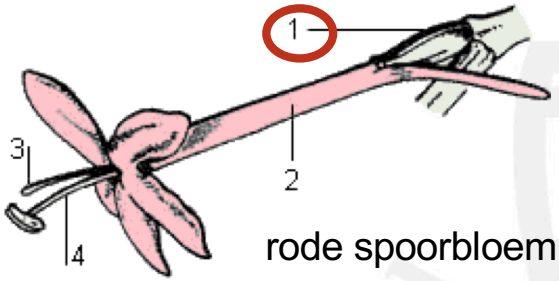
Find the sepal (kelkblad)



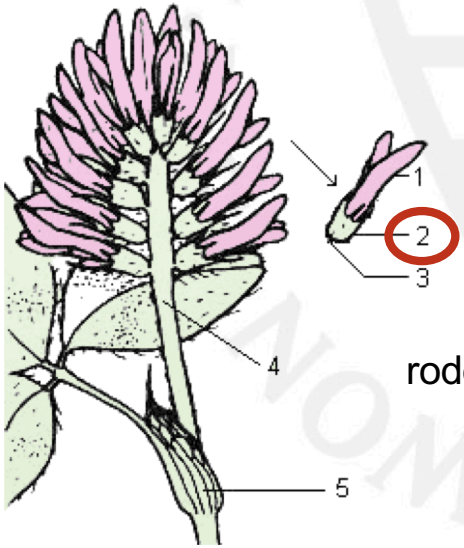
basilicum



liggend hertshooi



rode spoorbloem



rode klaver

Summary Statistics

- Summary statistics are numbers that summarize properties of the data
- Summarized properties include frequency, location and spread
  - Examples: location - mean
    spread - standard deviation
- Most summary statistics can be calculated in a single pass through the data

Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the data set
  - For example, given the attribute ‘gender’ and a representative population of people, the gender ‘female’ occurs about 50% of the time
- The mode of an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

Percentiles

- For continuous data, the notion of a percentile is more useful
- Given an ordinal or continuous attribute x and a number (percentage) p between 0 and 100, the pth percentile is a value $x_{p\%}$ such that $p\%$ of the observed values of x are less than $x_{p\%}$
- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of the observed values of x are less than $x_{50\%}$
- Special cases: median $x_{50\%}$ and quartiles $x_{25\%}$ and $x_{75\%}$

Measures of Location

- The mean is the most common measure of the location of a set of points:

$$\text{mean}(x) = \frac{1}{n}\sum_{k=1}^{n} x_k$$

- However, the mean is very sensitive to outliers
- Thus, the median or a trimmed mean is also commonly used:

$$\text{median}(x) = \begin{cases} x_{(r+1)}, & n = 2r+1 \quad (odd) \\ \frac{1}{2}\left(x_{(r)} + x_{(r+1)}\right), & n = 2r \quad (even) \end{cases}$$

Measures of Spread: Range and Variance

- Range is the difference between the max and min
- The variance or standard deviation is the most common measure of the spread of a set of points:

$$\text{variance}(x) = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \bar{x}\right)^2 \quad = \quad \text{standard-deviation}(x)^2$$

- Both are sensitive to outliers, so that other measures are often used:

$$\text{average-absolute-deviation}(x) = \frac{1}{n}\sum_{k=1}^{n}\left|x_k - \bar{x}\right|$$

$$\text{median-absolute-deviation}(x) = \text{median}\left(\left\{\left|x_1 - \bar{x}\right|,\ldots,\left|x_n - \bar{x}\right|\right\}\right)$$

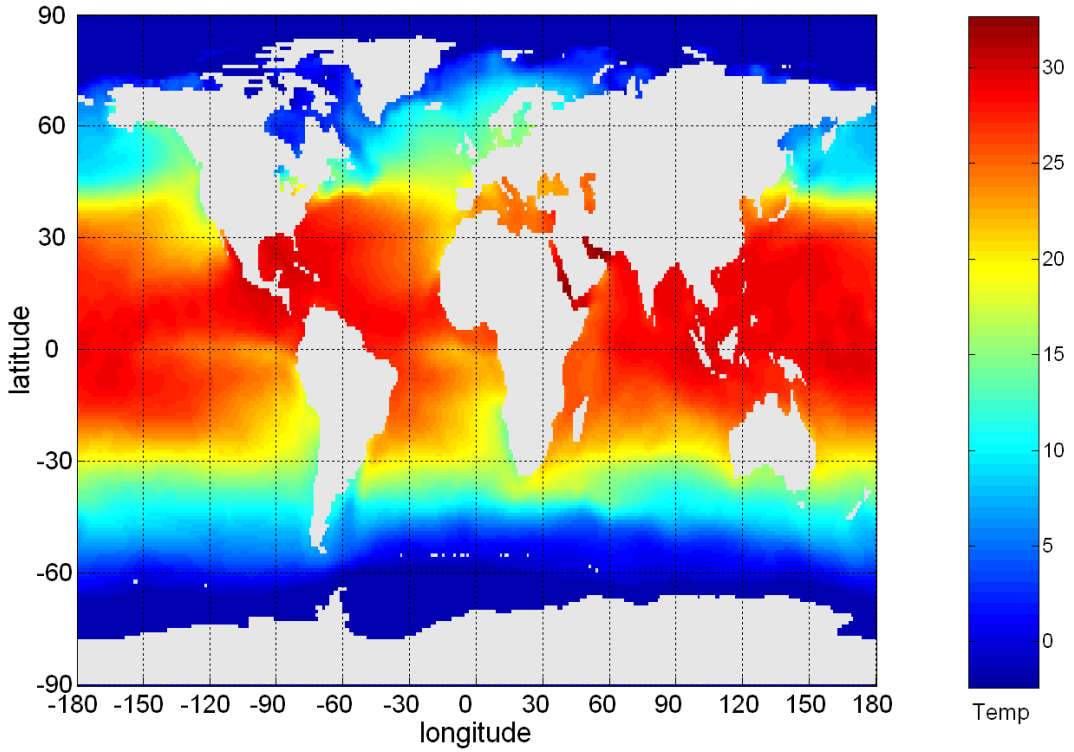
$$\text{interquartile-range}(x) = x_{75\%} - x_{25\%}$$

Visualization

- Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.
- Visualization of data is one of the most powerful and appealing techniques for data exploration
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

Example: Sea Surface Temperature

- The following shows the Sea Surface Temperature (SST) for July 1982
  - Tens of thousands of data points are summarized in a single figure



Representation

- Representation refers to the mapping of information to a visual format
- Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors
- Example:
  - Objects are often represented as points
  - Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
  - If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived.

Arrangement

- Is the placement of visual elements within a display
- Can make a large difference in how easy it is to understand the data

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

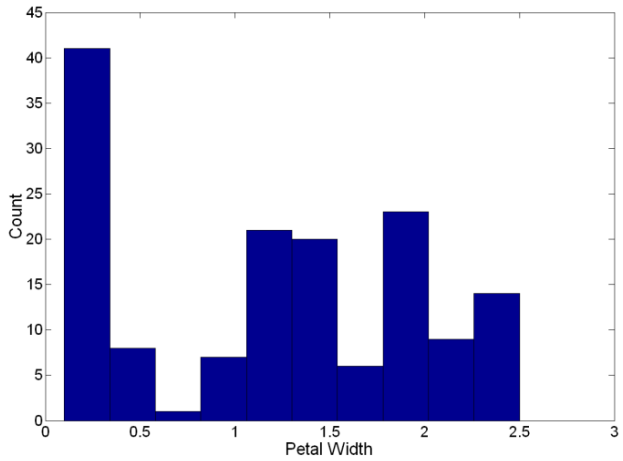
	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

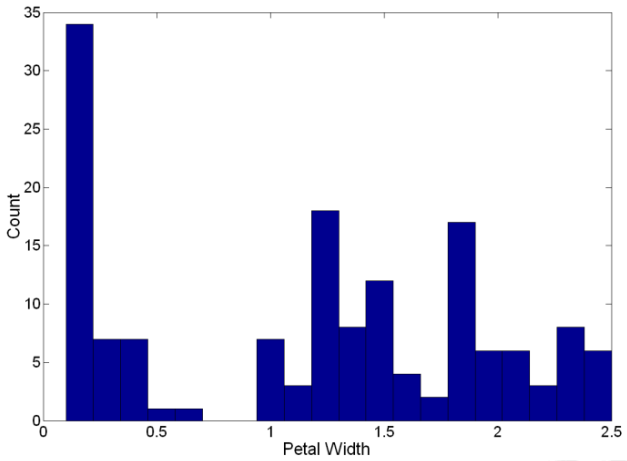
Selection

- Selection refers to the elimination or the de-emphasis of certain objects and attributes
- Selection may involve choosing a subset of attributes
  - Dimensionality reduction is often used to reduce the number of dimensions to two or three
  - Alternatively, pairs of attributes can be considered
- Selection may also involve choosing a subset of objects
  - A region of the screen can only show so many points
  - Can sample, but want to preserve points in sparse areas

Visualization Techniques: Histograms

- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)





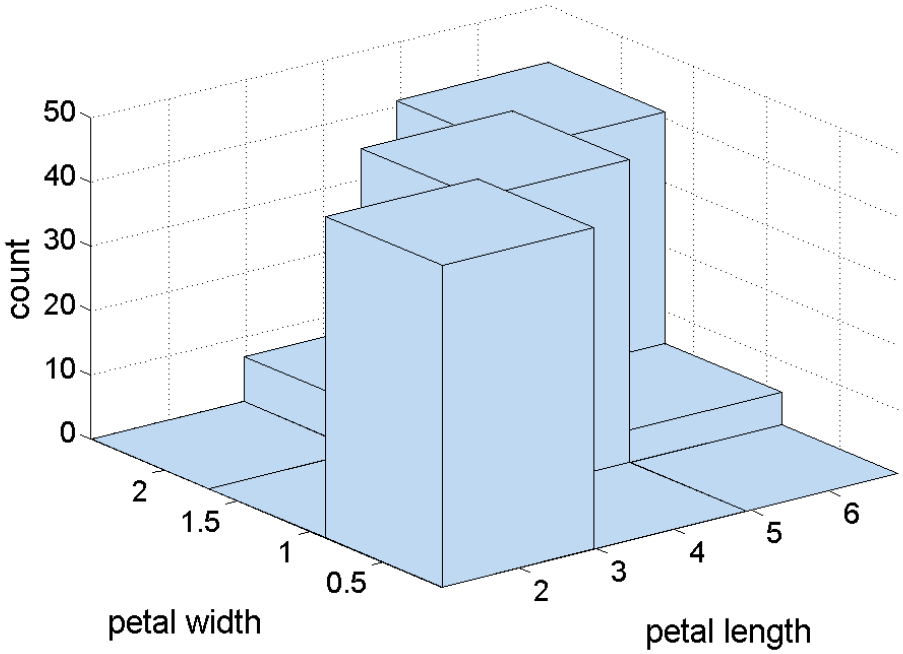
Python Code for Histogram

```
% Load iris data set

f = loadmat(“fisheriris.mat”)    % meas., species (classes)
meas = f[‘meas’]
species = f[‘species’]

% Plot histogram for 10 bins

plt.subplot(2,2,1)
plt.hist(meas[:,3], bins=10)
plt.xlabel('Petal Width'),
plt.ylabel('Count')
plt.title('Ten Bins')
```

Two-Dimensional Histograms

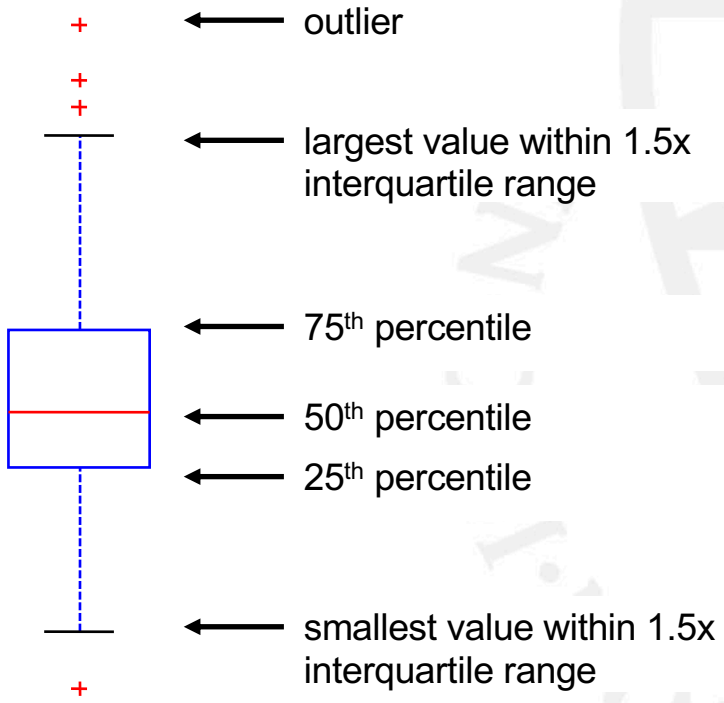
- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
- Code:

```
hist3(meas(:,3:4),[3,3])
xlabel('petal length')
ylabel('petal width')
zlabel('count')
```



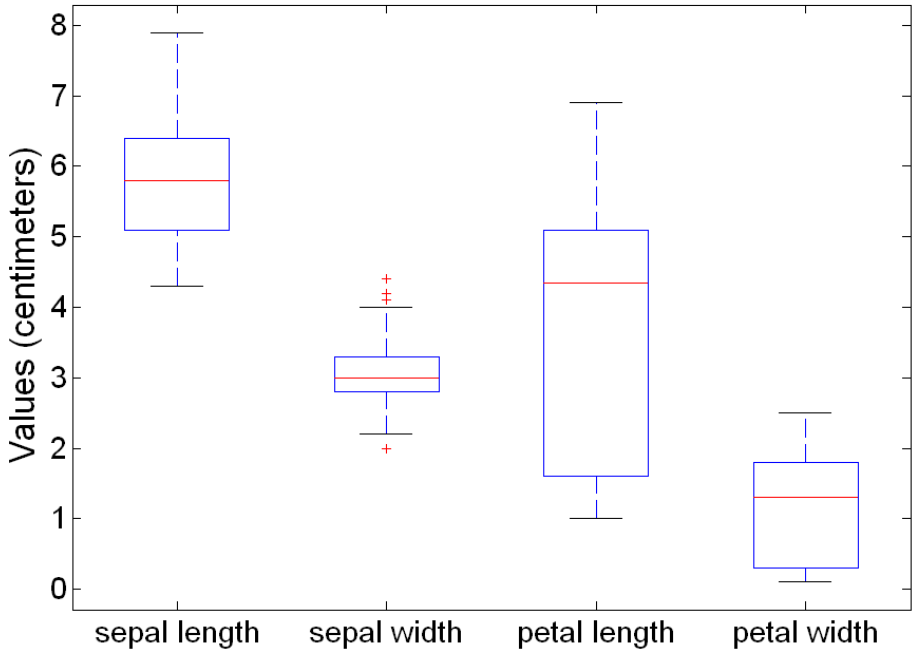
Visualization Techniques: Box Plots

- Invented by J. Tukey
- Another way of displaying the distribution of data

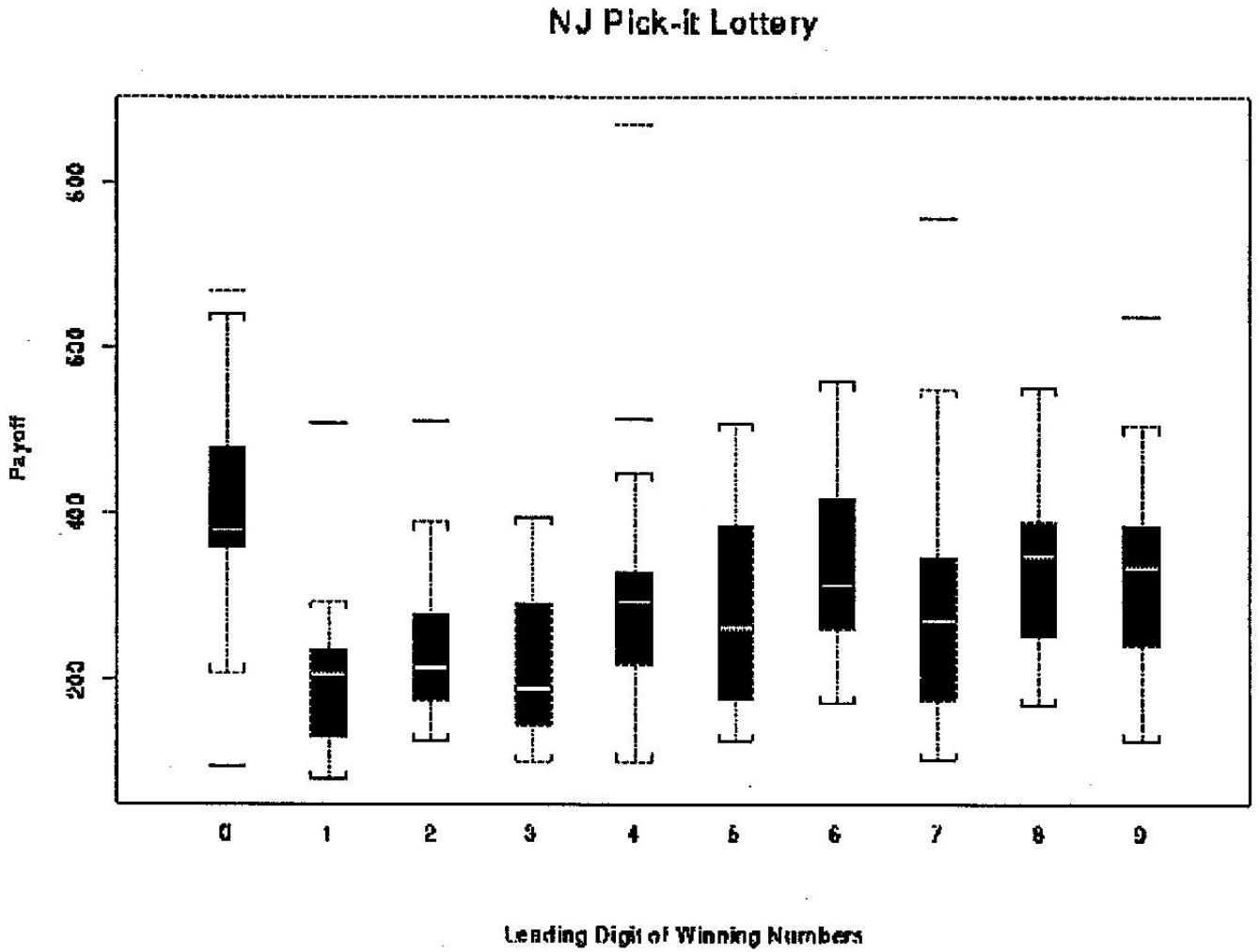


Box Plots in Python

```
plt.boxplot(meas)
plt.ylabel('Values (centimeters)')
plt.xticks([1, 2, 3, 4],
           ['sepal length','sepal width',
            'petal length','petal width'])
```



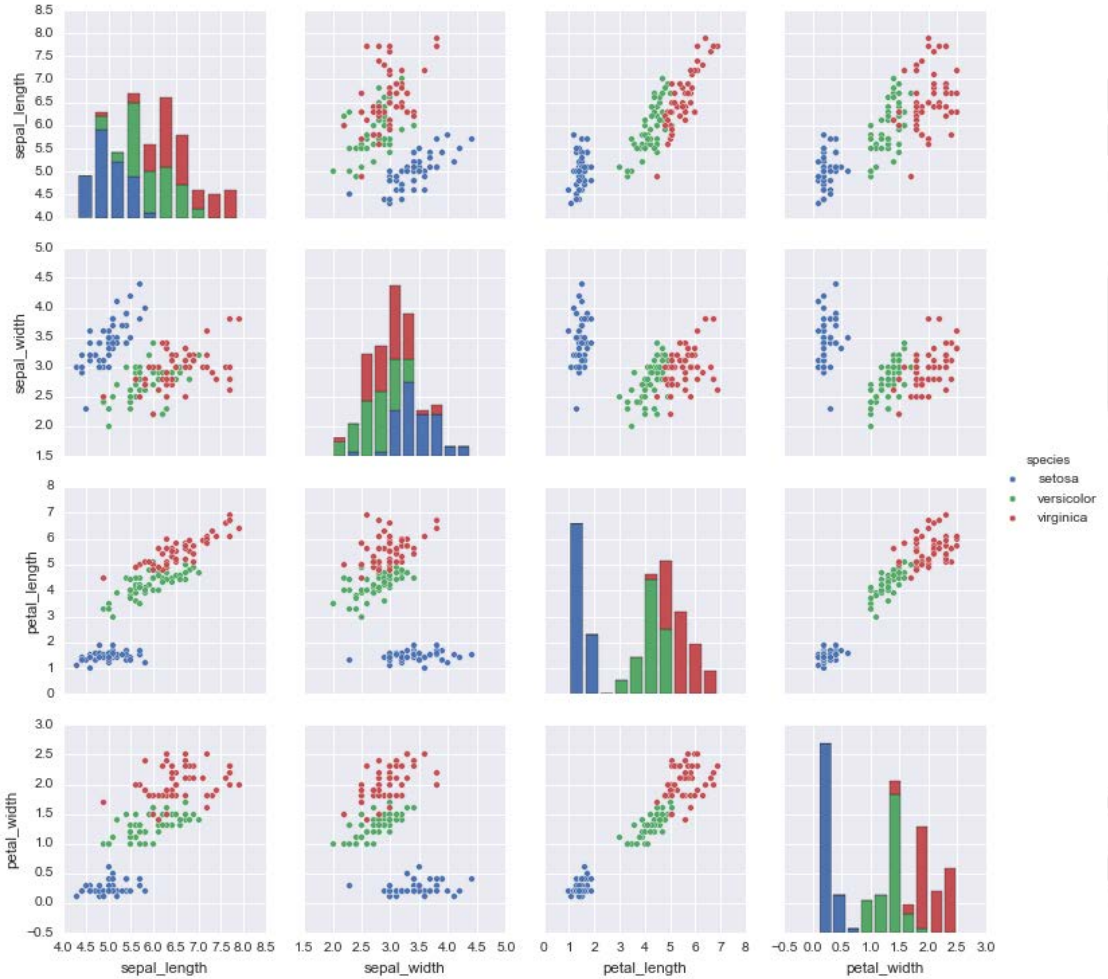
Boxplot of Lottery Pay-off



Visualization Techniques: Scatter Plots

- Attributes values determine the position
- Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
- Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
- Arrays of scatter plots can compactly summarize the relationships of several pairs of attributes

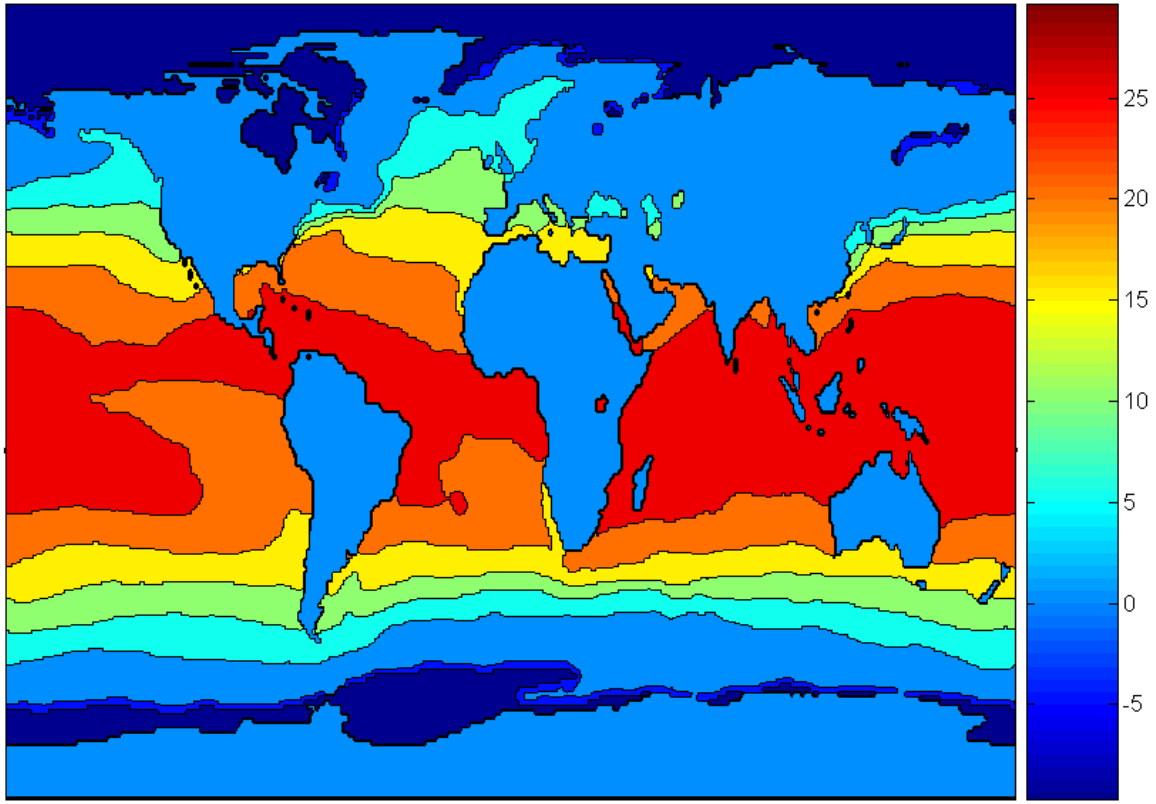
Python: use pandas or seaborn



Visualization Techniques: Contour Plots

- Useful when a continuous attribute is measured on a spatial grid
- They partition the plane into regions of similar values
- The contour lines that form the boundaries of these regions connect points with equal values
- The most common example is contour maps of elevation
- Can also display temperature, rainfall, air pressure, etc.

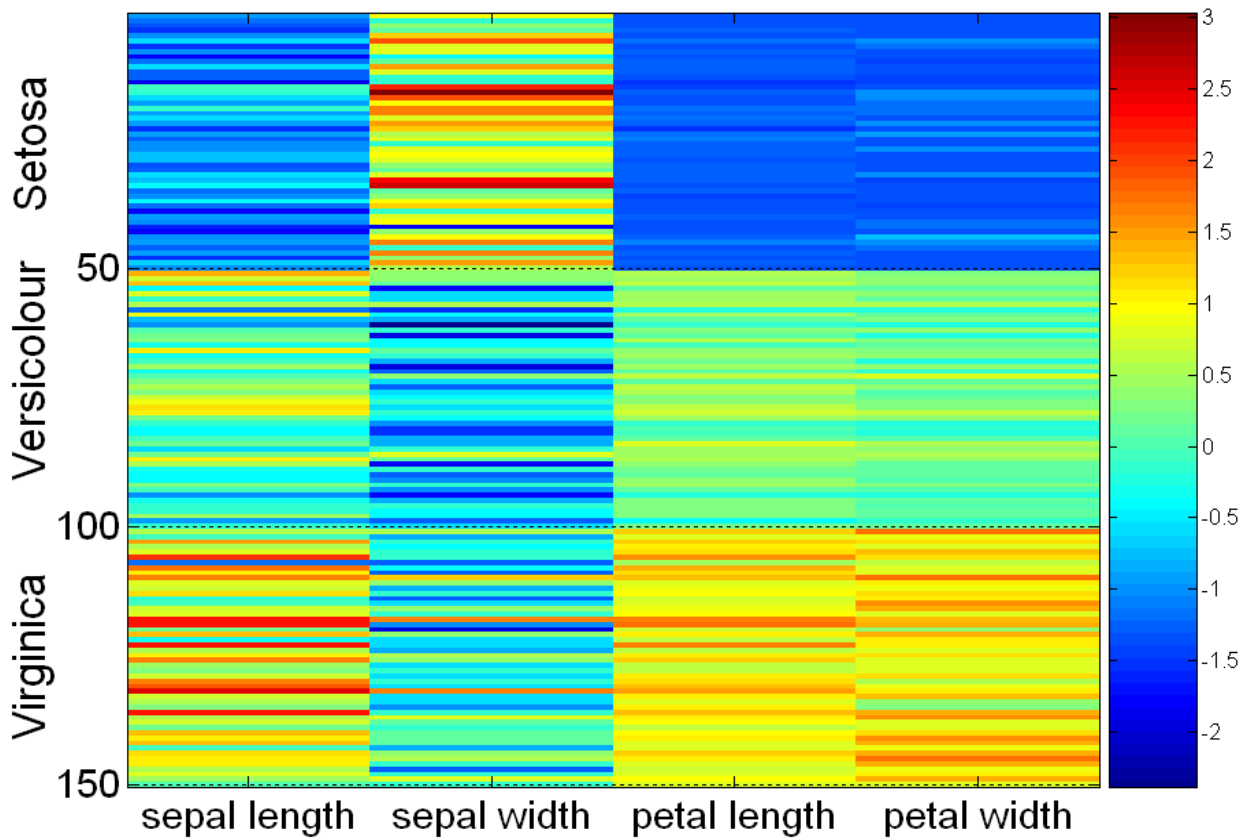
Contour Plot Example: SST Dec, 1998



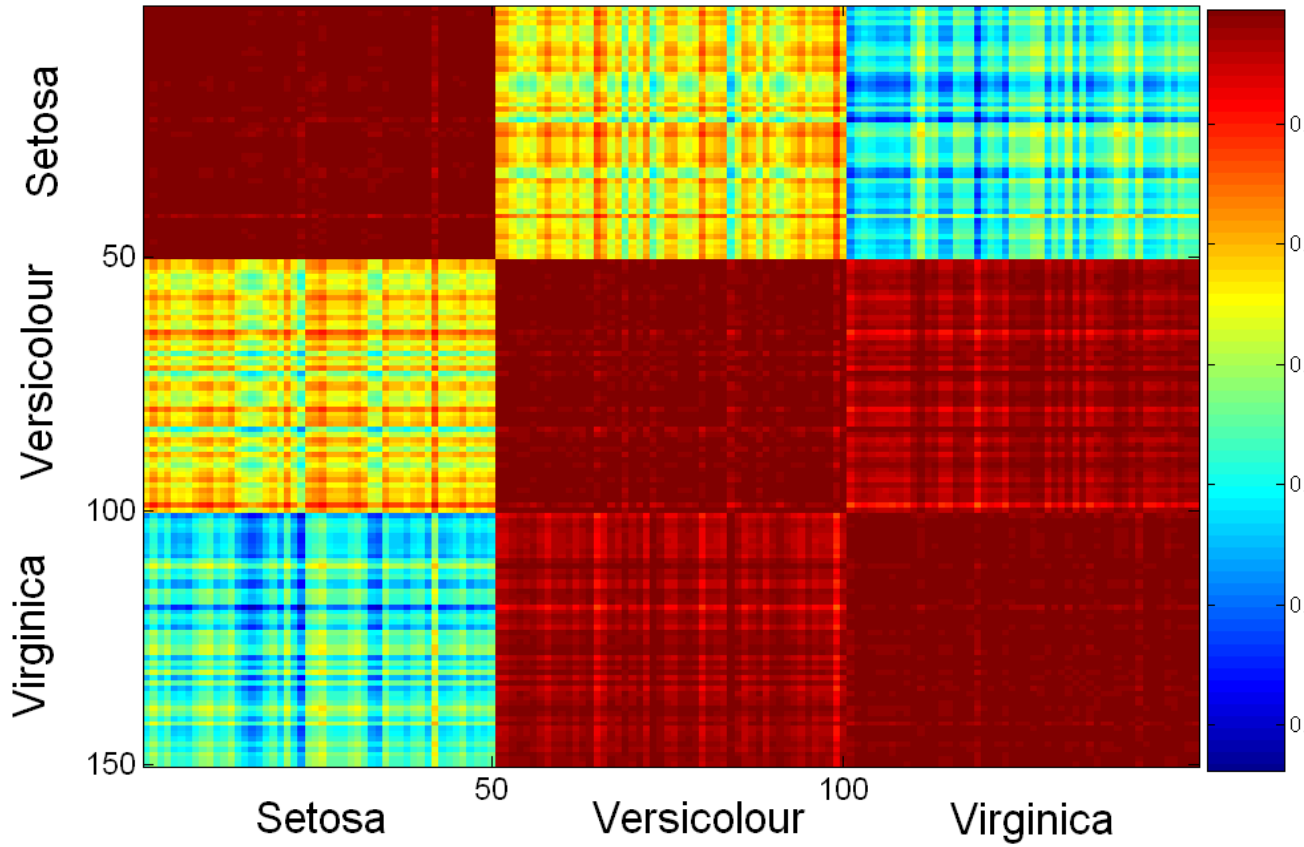
Visualization Techniques: Matrix Plots

- Can plot the data matrix
- This can be useful when objects are sorted according to class
- Typically, the attributes are normalized to prevent one attribute from dominating the plot
- Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects

Visualization of the Iris Data Matrix



Visualization of the Iris Correlation Matrix

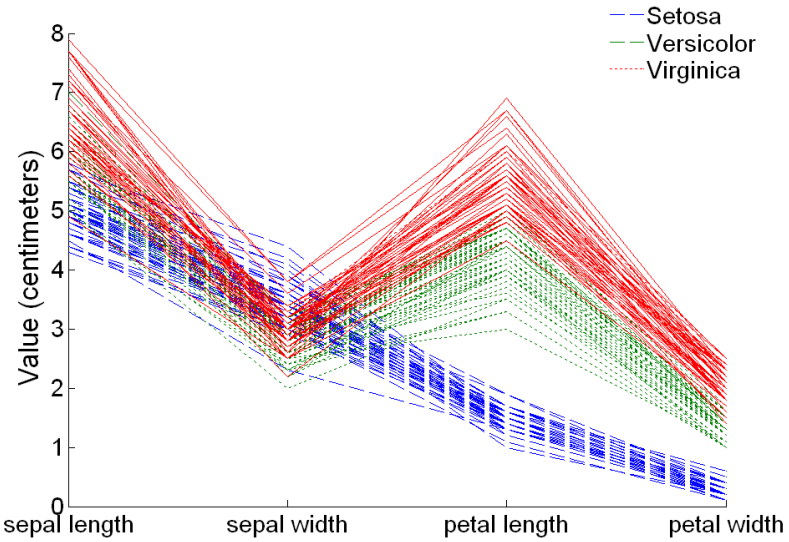


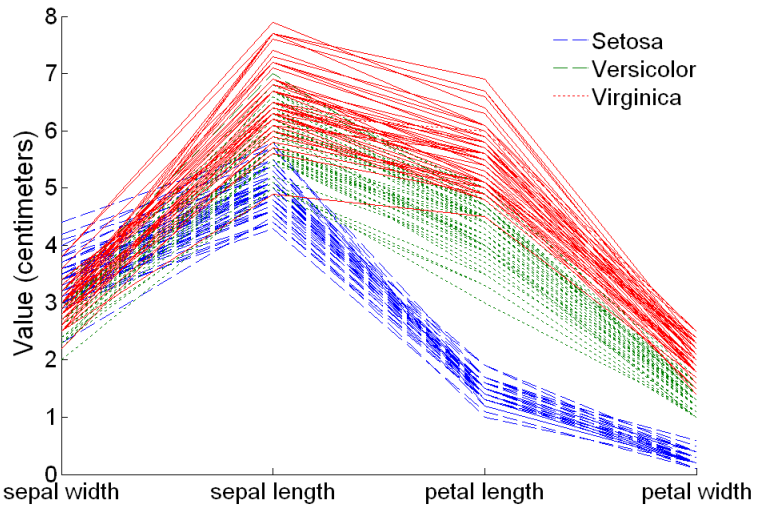
Visualization Techniques: Parallel Coordinates

- Used to plot the attribute values of high-dimensional data
- Instead of using perpendicular axes, use a set of parallel axes
- The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
- Thus, each object is represented as a line
- Often, the lines representing a distinct class of objects group together, at least for some attributes
- Ordering of attributes is important in seeing such groupings

Parallel Coordinates Plots for Iris Data

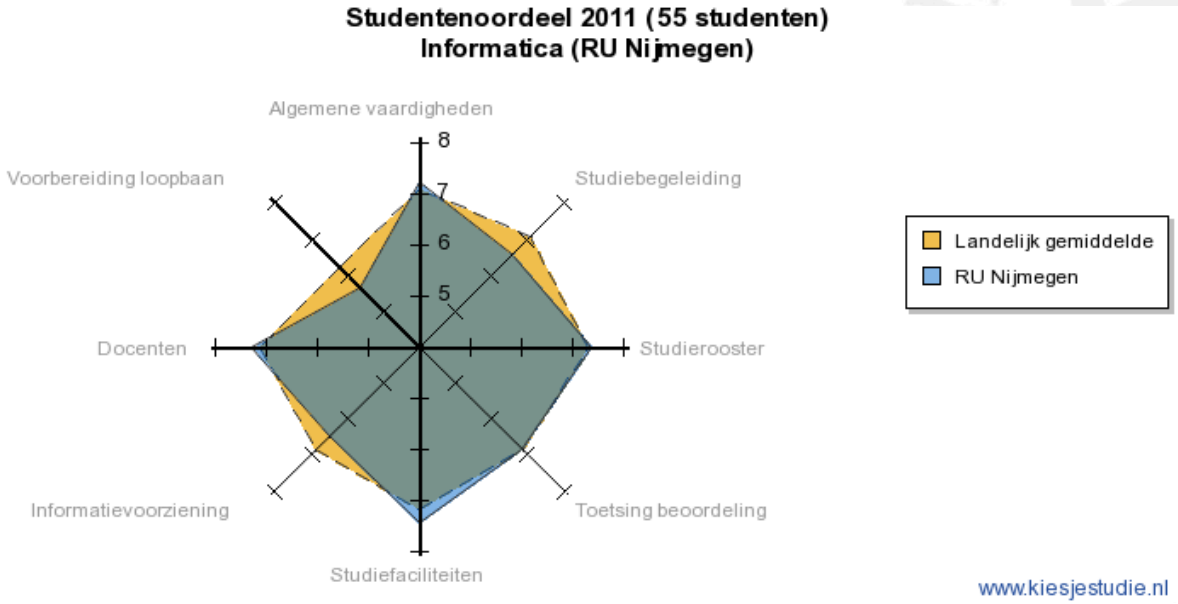
```
parallelcoords(meas,'group',species,'labels',varnames)
parallelcoords(meas(:,[2,1,3,4]),'group',species,...
    'labels',varnames([2,1,3,4]))
ylabel('Values (centimeters)')
```





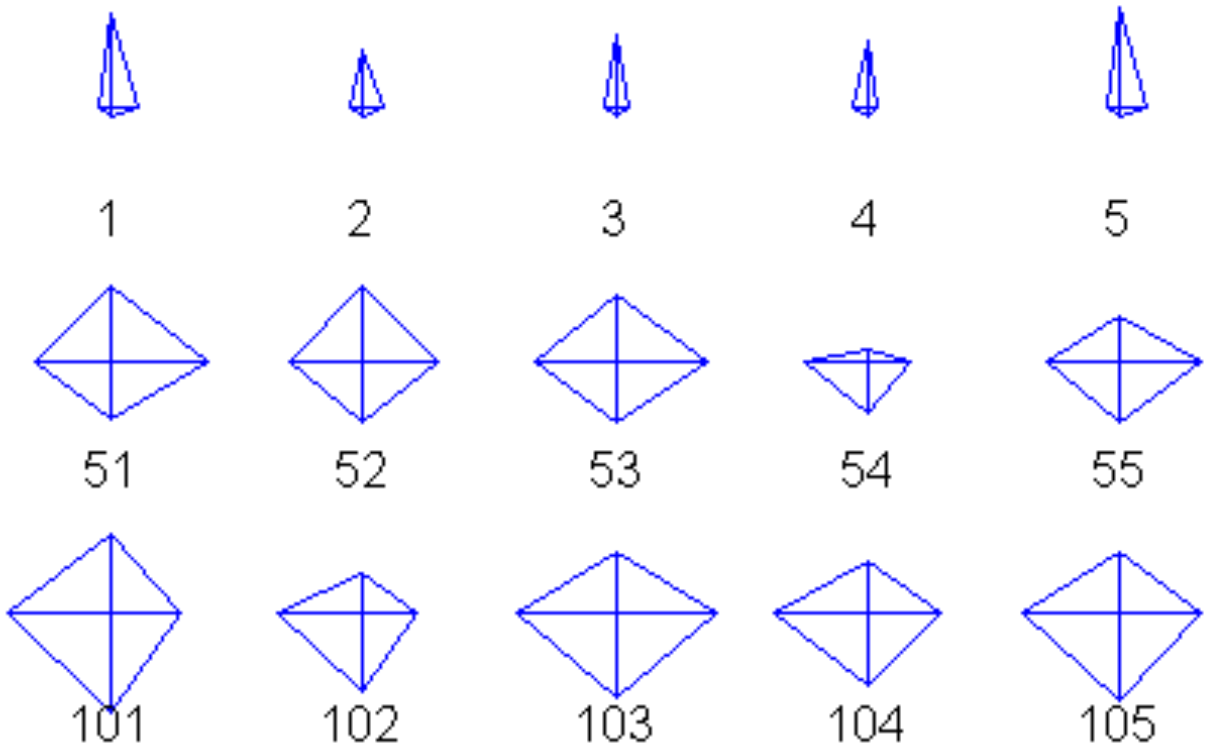
Star Plots

- Similar approach to parallel coordinates, but axes radiate from a central point
- The line connecting the values of an object is a polygon

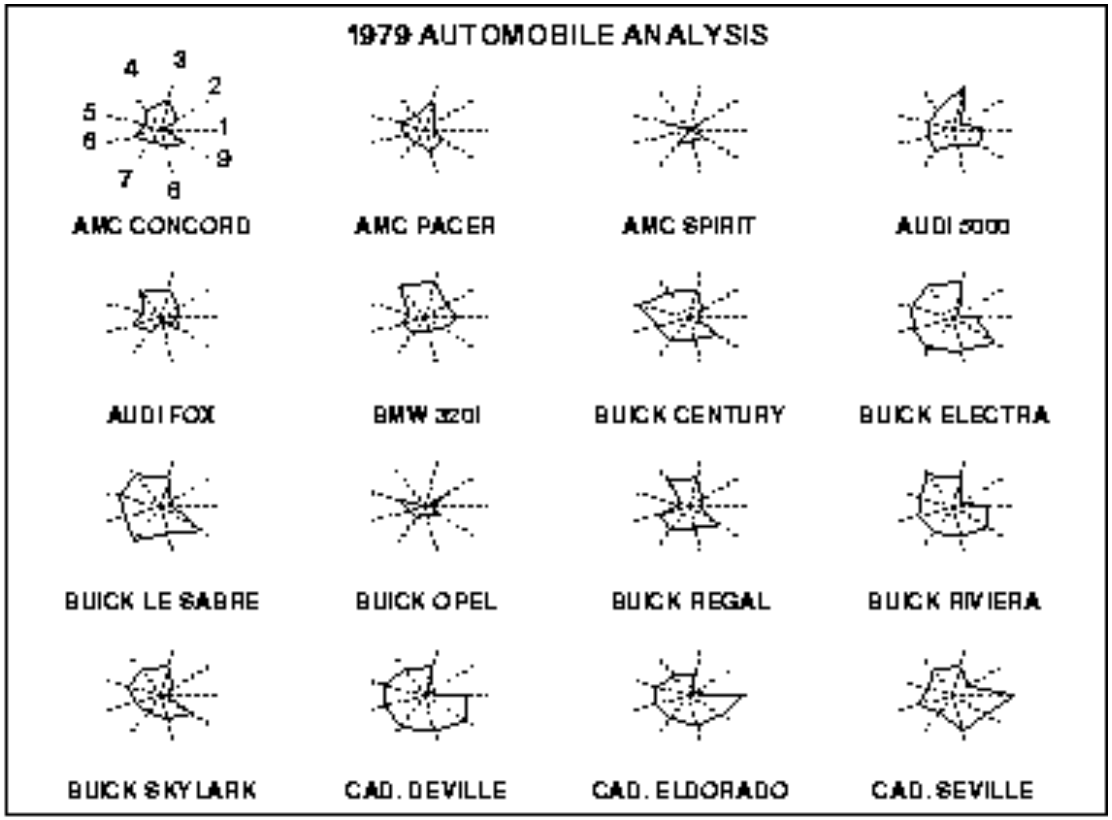


Star Plots for Iris Data

```
glyphplot(meas(indices,:),'ObsLabels',labels)
```



Star Plots for Cars



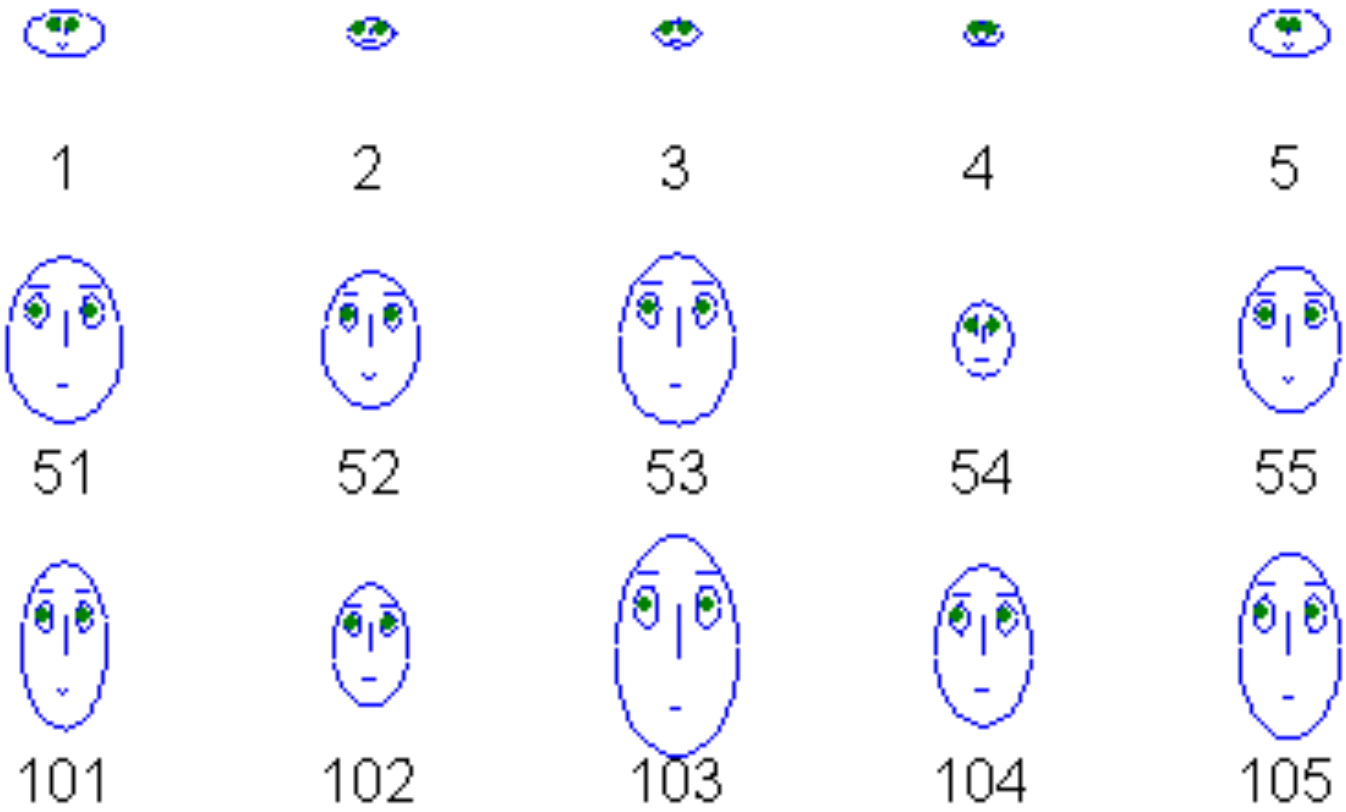
1. Price
2. Mileage (MPG)
3. 1978 Repair Record (1 = Worst, 5 = Best)
4. 1977 Repair Record (1 = Worst, 5 = Best)
5. Headroom
6. Rear Seat Room
7. Trunk Space
8. Weight
9. Length

Chernoff Faces

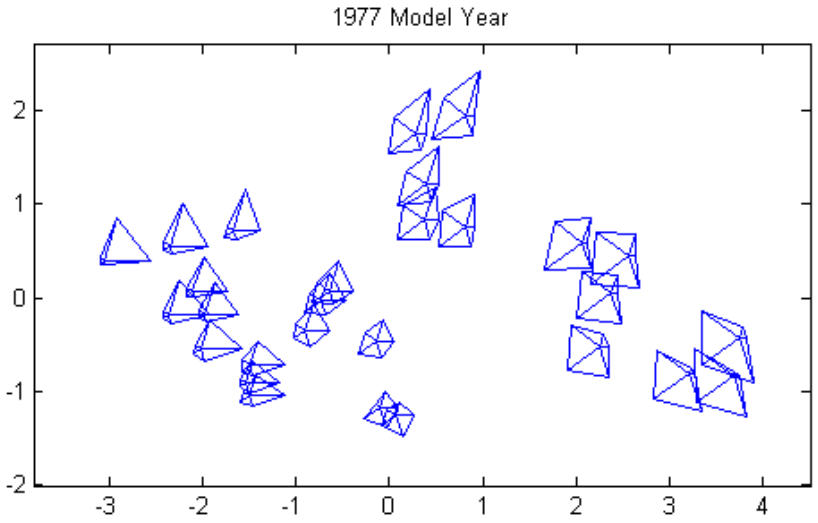
- Approach created by Herman Chernoff
- This approach associates each attribute with a characteristic of a face
- The values of each attribute determine the appearance of the corresponding facial characteristic
- Each object becomes a separate face
- Relies on human's ability to distinguish faces

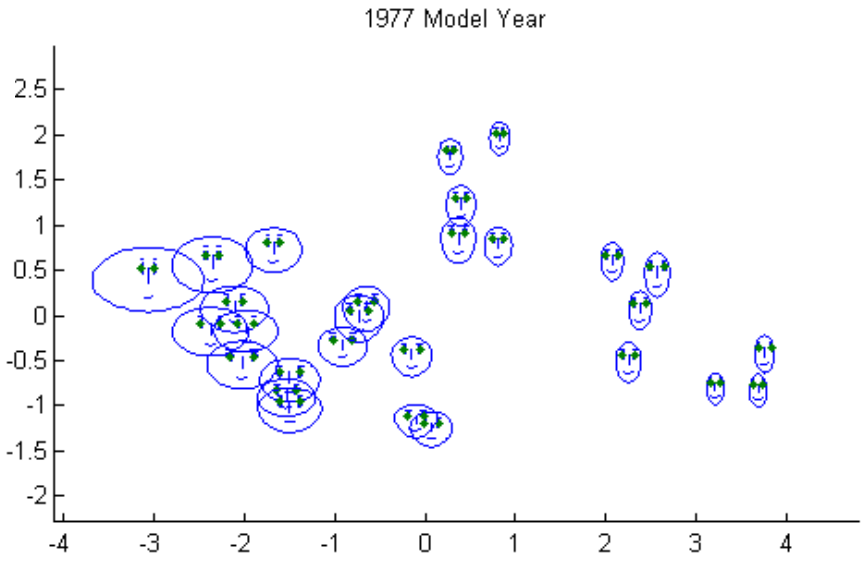
Chernoff Faces for Iris Data

```
glyphplot(meas(indices,:),'ObsLabels',labels,'Glyph','face')
```



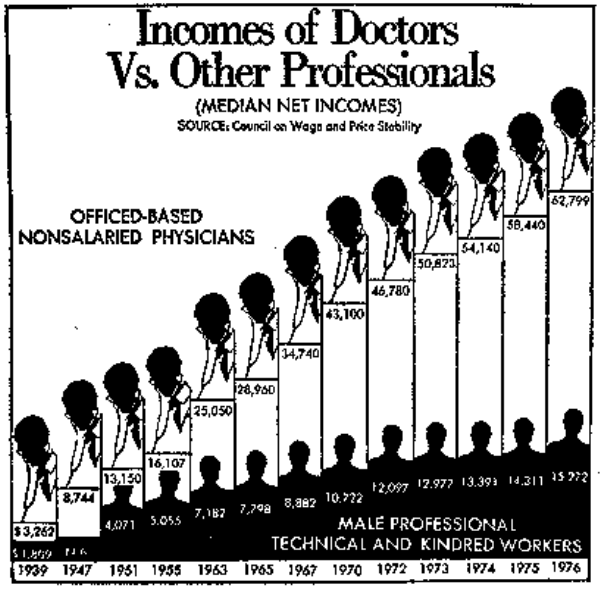
Clustering Combinations

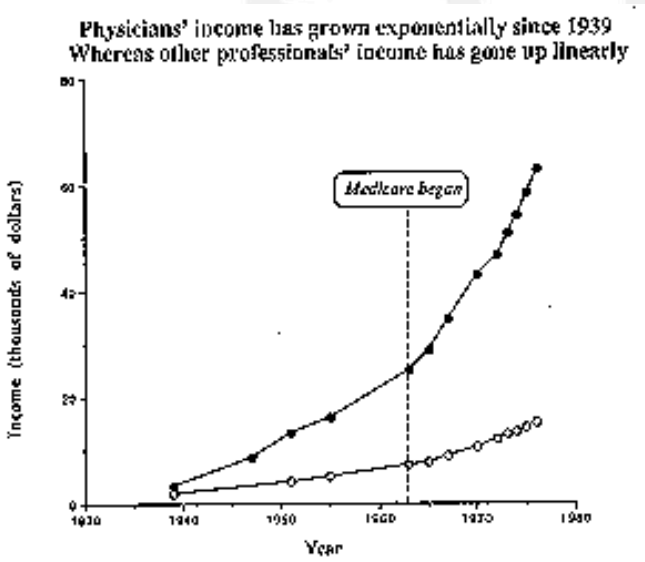




Cheating

- Graph showing change in income of doctors vs. other professionals
- Appears to indicate a more or less linear increase in both
- Cheating: axes are unevenly spaced!
- Correct spacing reveals exponential increase





Lie factor

- Tufte (1983): “The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the quantities represented”

- Deviation cast in terms of formula:

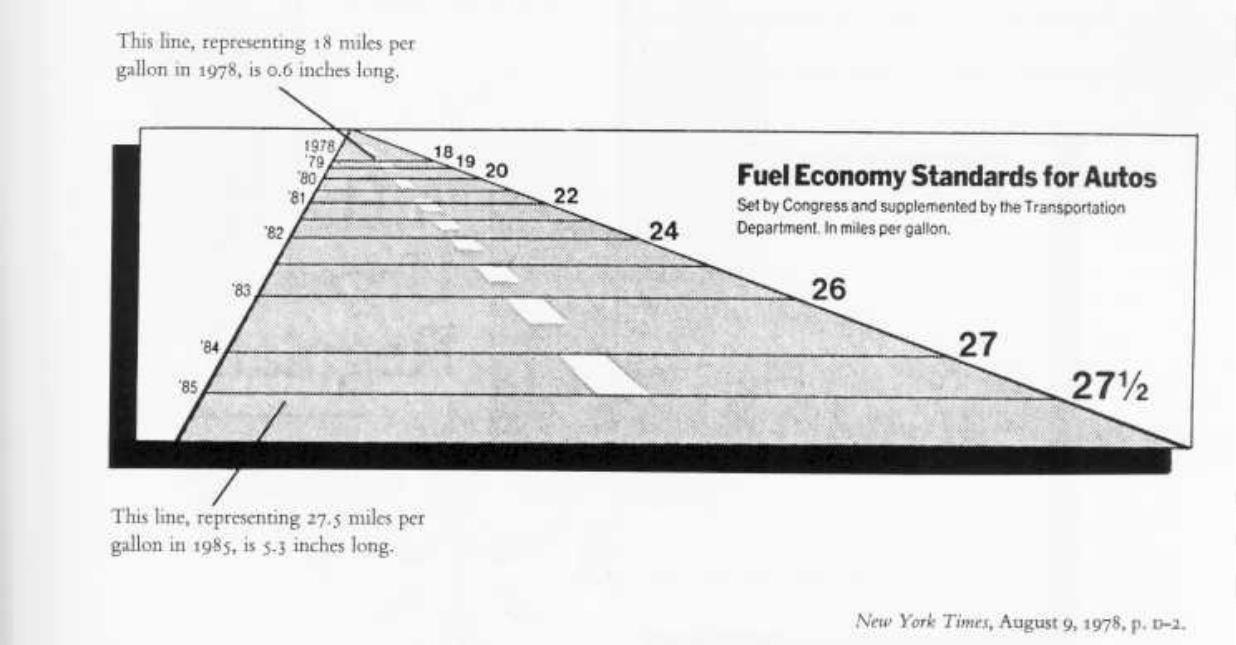
$$\text{Lie factor} = \frac{\text{size of effect shown in graph}}{\text{size of effect in data}}$$

- Where:

$$\text{size of effect} = \frac{\left|\text{second value - first value}\right|}{\text{first value}}$$

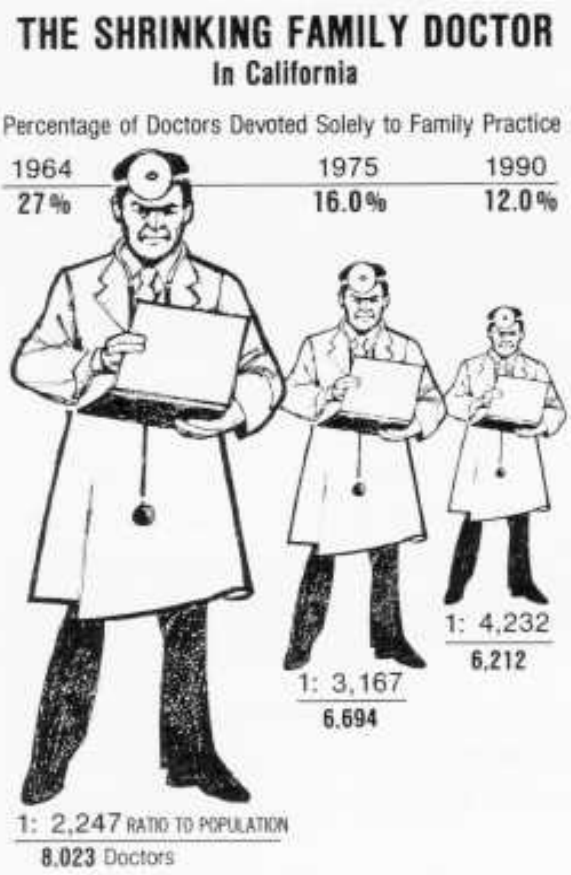
The Lie Factor (1)

- Mandated fuel economy standards set by the US Department of Transportation
- The standard required an 53% increase in mileage from 18 to 27.5
- The magnitude of increase shown in the graph is 783%
- Lie factor = (783/53) = 14.8!



The Lie Factor (2)

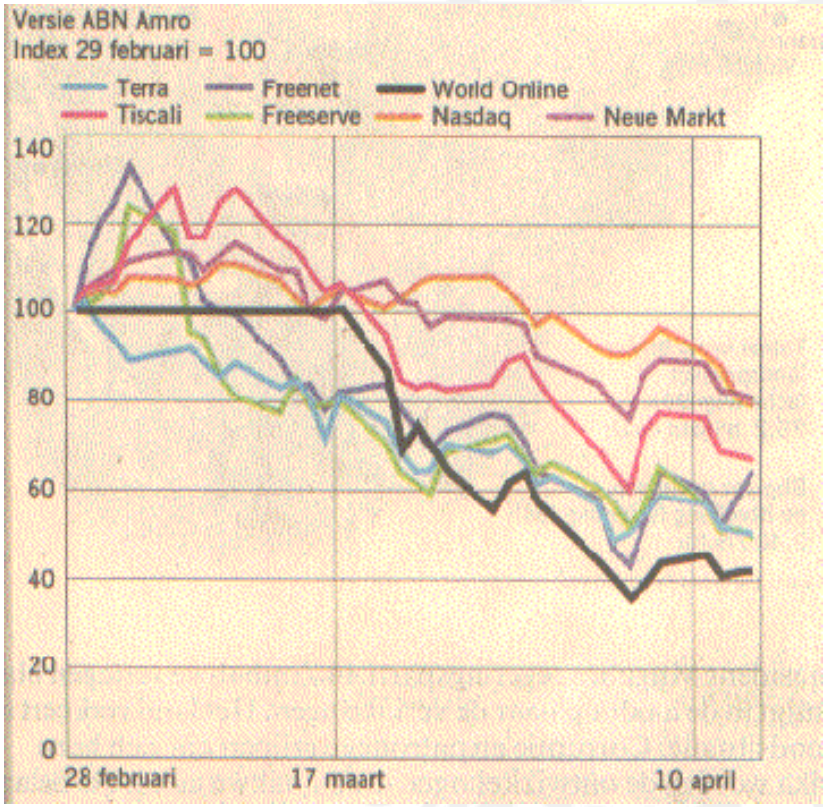
- Changes in the scale of the graphic should always correspond to changes in the data being represented
- This graph violates that principle by using area to show one-dimensional data
- Lie factor: 2.8



Los Angeles Times, August 5, 1979, p. 3.

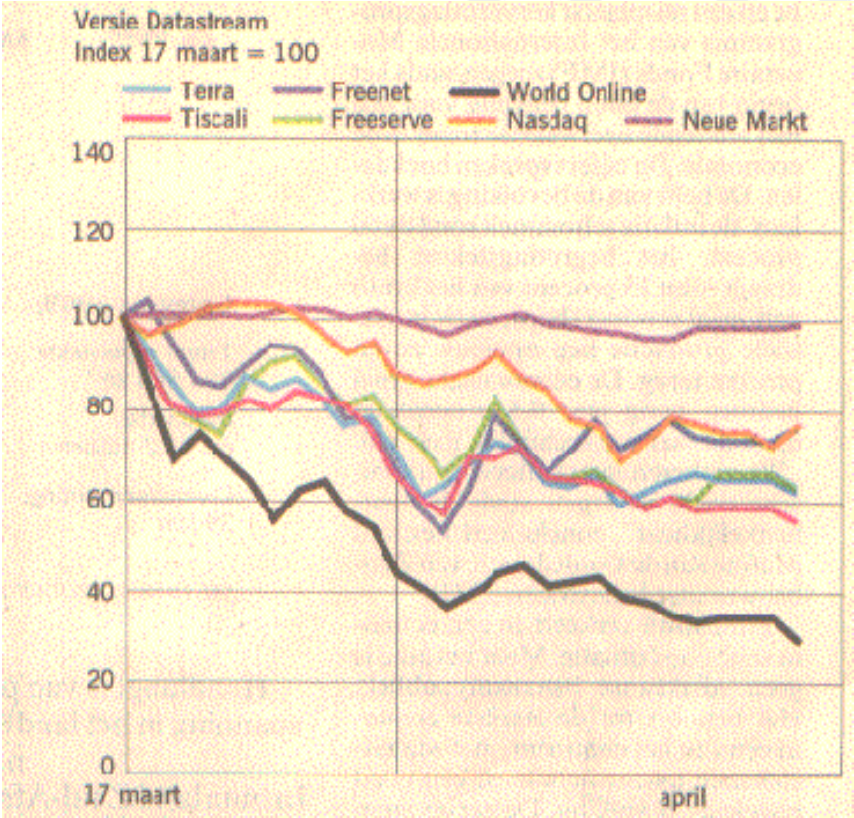
World On Line (1)

- Stock prices of the Dutch internet provider World On Line (WOL) halved within less than two weeks after entering the Amsterdam Stock Exchange
- ABN AMRO: “We could not foresee this. Many other funds were in a downfall too; some of them more than WOL”
- The bank illustrated this by the graph on the right.



World On Line (2)

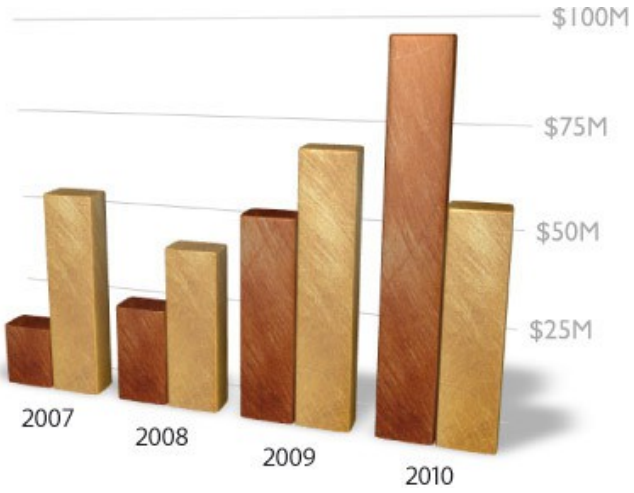
- No reason for a flat line between February 28 and March 17, the start WOL's stock market quotation.
- Unshifting the base date for the index numbers to March 17 give this graph, in which WOL appears to be very quickly heading down the toilet…

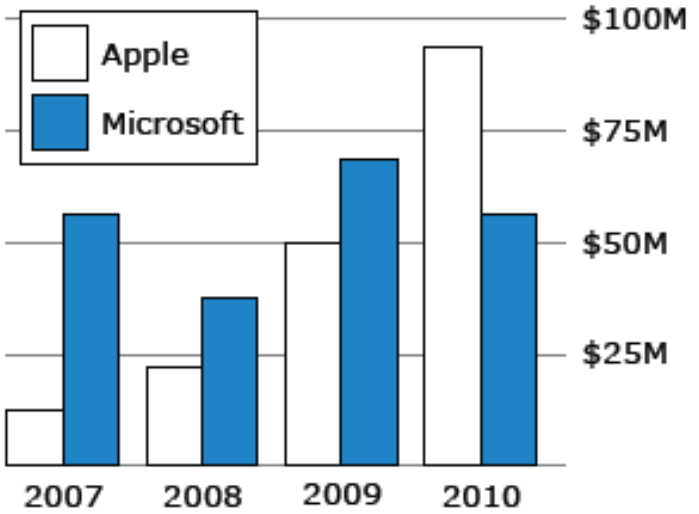


ACCENT

- Apprehension
  - Is it easy to see what is important in the graph?
- Clarity
  - Are the most important elements visually most prominent?
- Consistency
  - Have you used the same colors, shapes, etc. as in other graphs?
- Efficiency
  - Does it convey its information in the most simple and efficient way?
- Necessity
  - Are all elements of the graph necessary to represent data?
- Truthfulness
  - Does the graph represent the data correctly?

Apple vs Microsoft

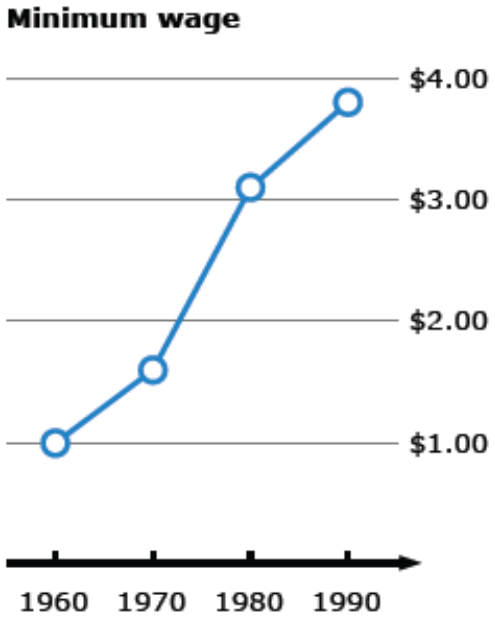




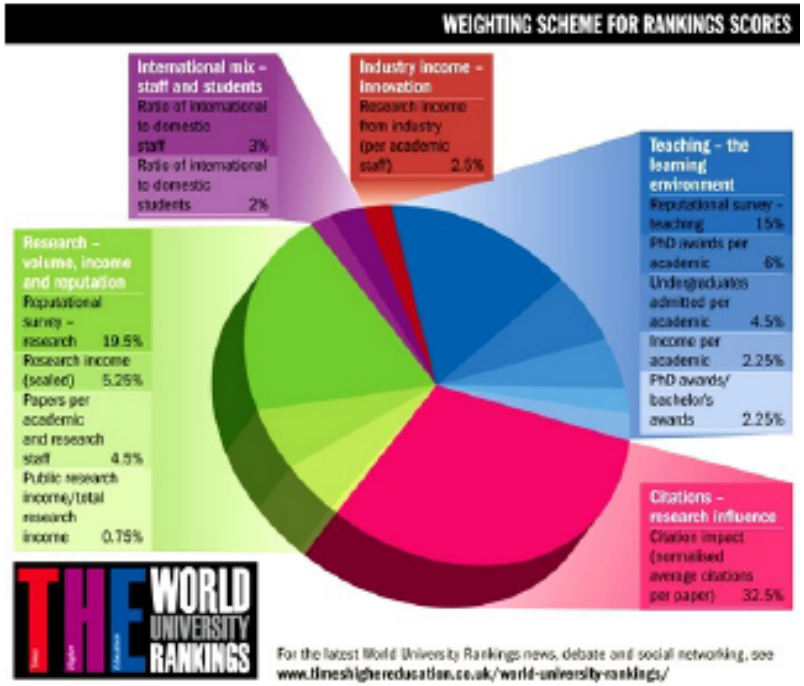
Minimum wage

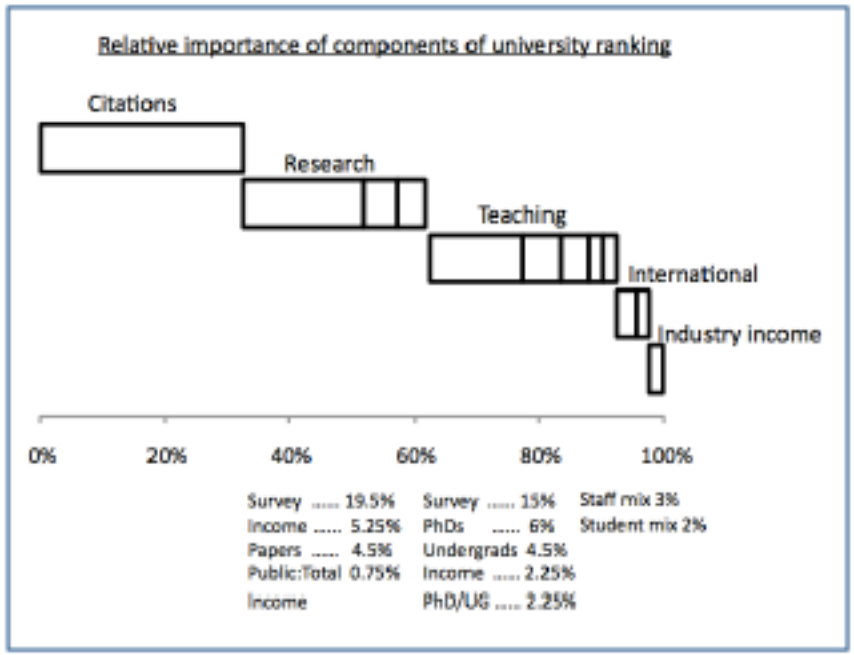
Minimum wage

1960		$1.00
1970		$1.60
1980		$3.10
1990		$3.80

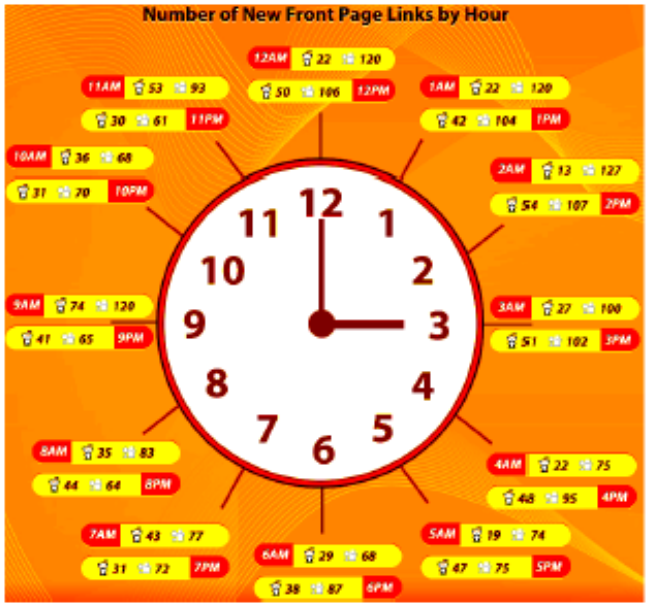


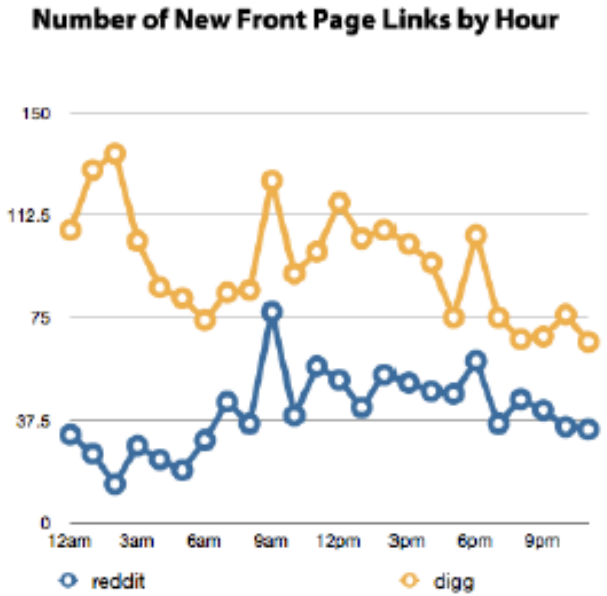
University Rankings





New Links


Number of New Front Page Links by Hour
12AM 22 120
50 106 12PM
1AM 22 120
42 104 1PM
2AM 13 127
54 107 2PM
3AM 27 100
51 102 3PM
4AM 22 75
48 95 4PM
5AM 19 74
47 75 5PM
6AM 29 68
38 87 6PM
7AM 43 77
31 72 7PM
8AM 35 83
44 64 8PM
9AM 74 120
41 65 9PM
10AM 36 68
31 70 10PM
11AM 53 93
30 61 11PM
12
1
2
3
4
5
6
7
8
9
10
11


Number of New Front Page Links by Hour
150
112.5
75
37.5
0
12am
3am
6am
9am
12pm
3pm
6pm
9pm
reddit
digg

Life Expectancy