

Data Mining: Data Exploration

Tom Heskes

What is data exploration?

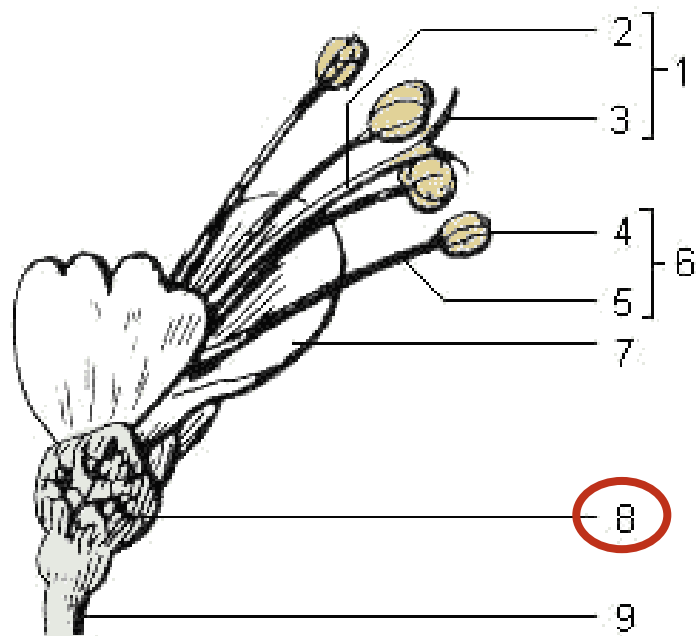
- A preliminary exploration of the data to better understand its characteristics
- Key motivations of data exploration include
 - Helping to select the right tool for preprocessing or analysis
 - Making use of humans' abilities to recognize patterns
- Related to the area of Exploratory Data Analysis (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>

Iris Sample Data Set

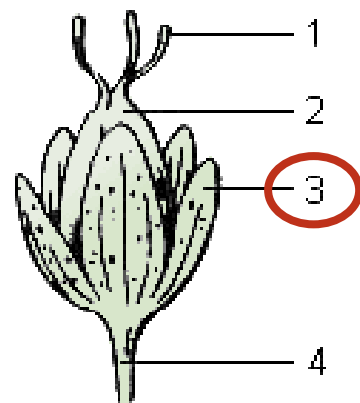
- Many of the exploratory data techniques are illustrated with the Iris Plant data set
- Can be obtained from the UCI Machine Learning Repository
<http://www.ics.uci.edu/~mlearn/MLRepository.html>
- From the statistician Douglas Fisher
- Three flower types (classes):
 - Setosa
 - Virginica
 - Versicolour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



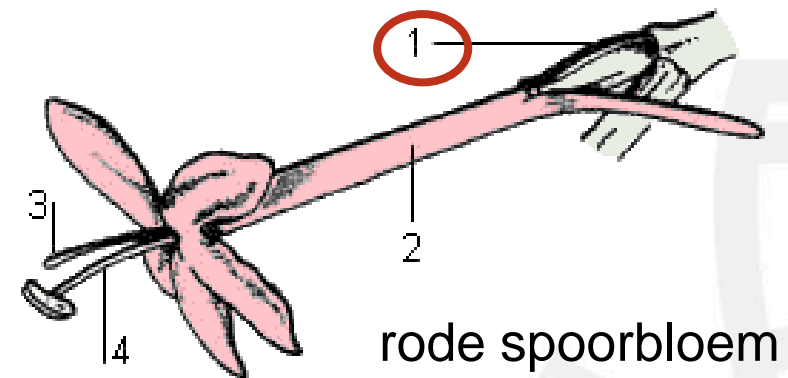
Find the sepal (kelkblad)



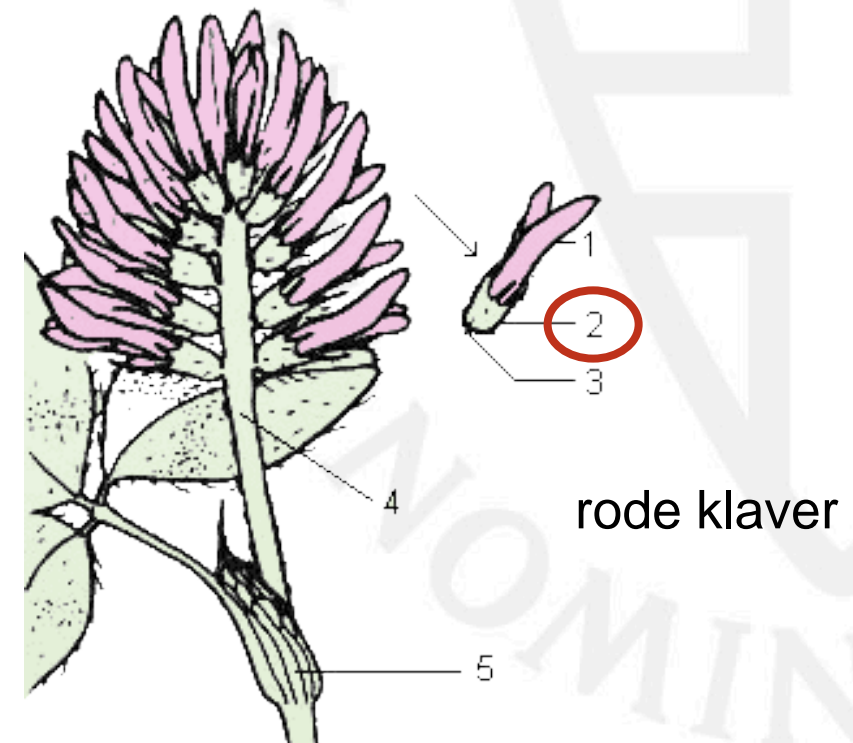
basilicum



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 p th 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:

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

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:

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

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

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

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

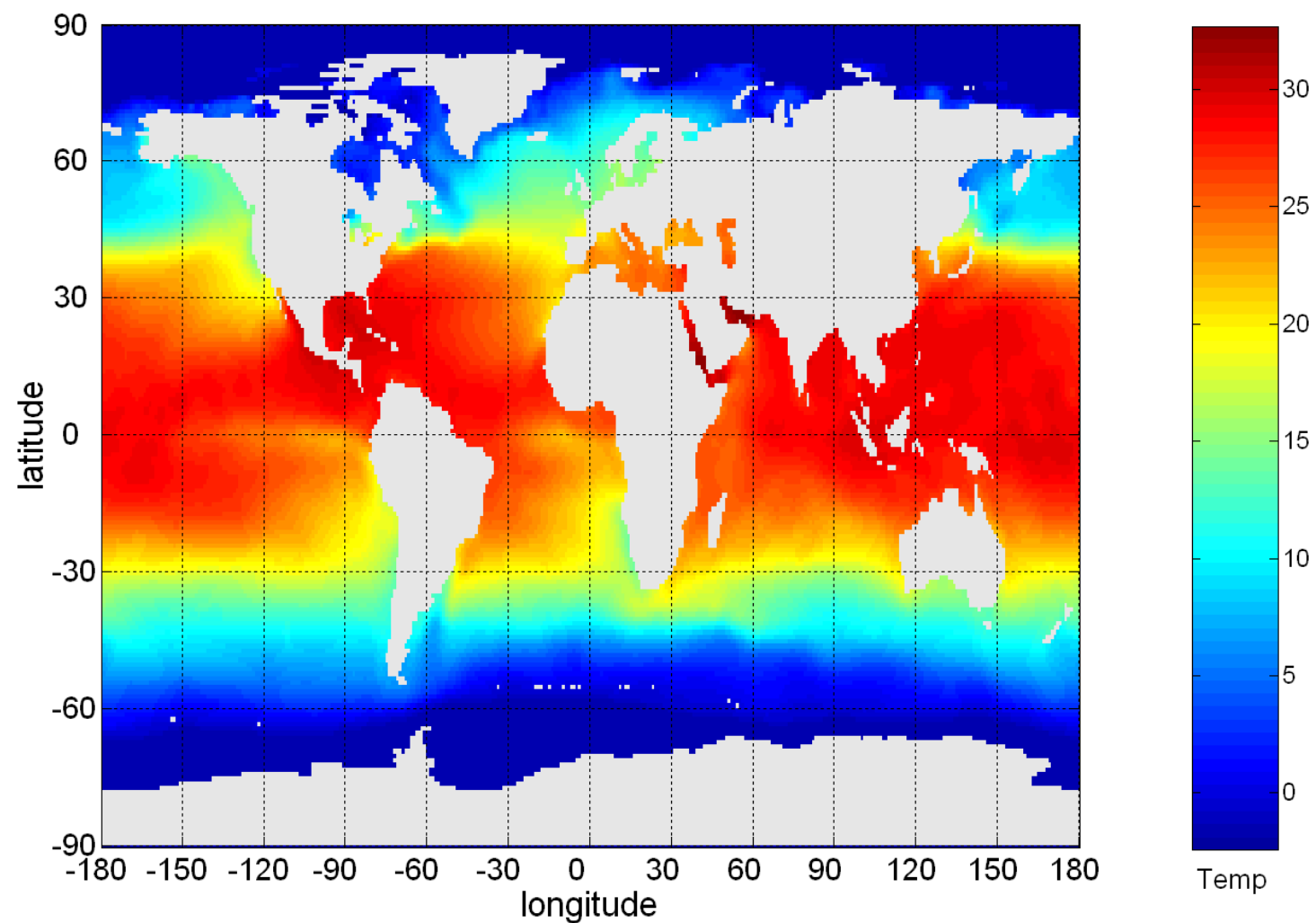
$$\text{interquartile-range}(x) = x_{75\%} - x_{25\%}$$

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

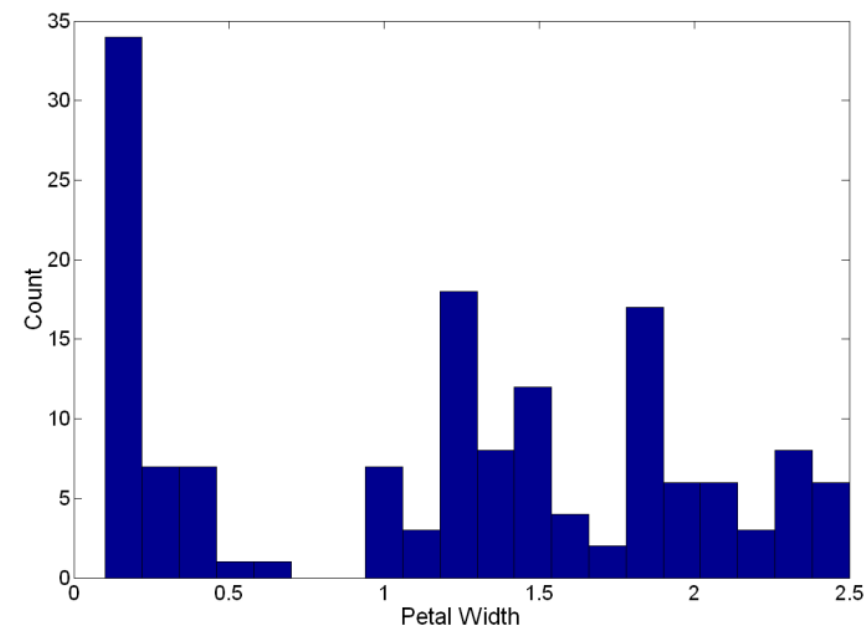
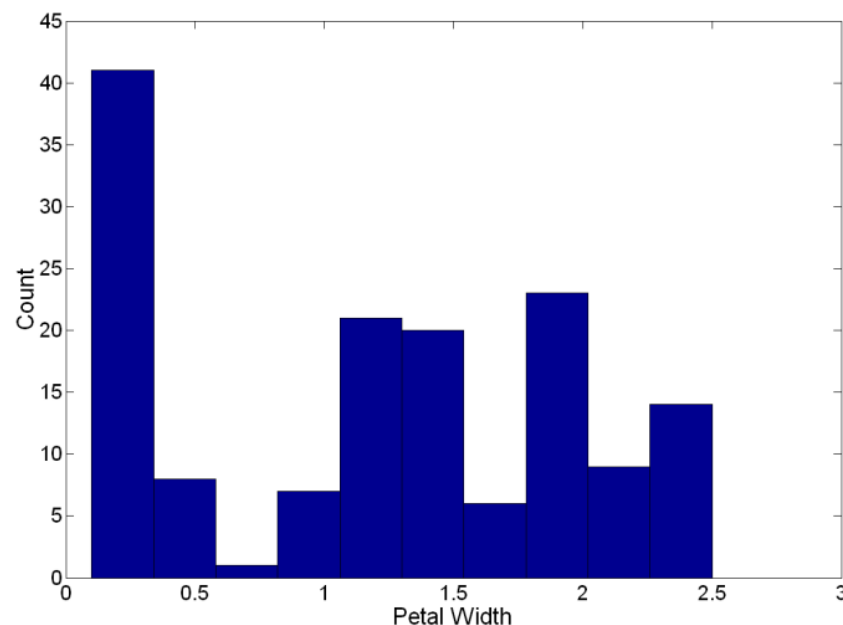
	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

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)



MATLAB Code for Histogram

```
% Load iris data set

load fisheriris    % meas (measurements), species (classes)

% Plot histograms

subplot(221)
hist(meas(:,4),10)
xlabel('Petal Width'), ylabel('Count')
title('Ten Bins')

subplot(222)
hist(meas(:,4),20)
xlabel('Petal Width'), ylabel('Count')
title('Twenty Bins')
```


Python Code for Histogram

```
# Load iris data set

f = loadmat("fisheriris.mat") % meas, species
meas = f['meas']
species = f['species']

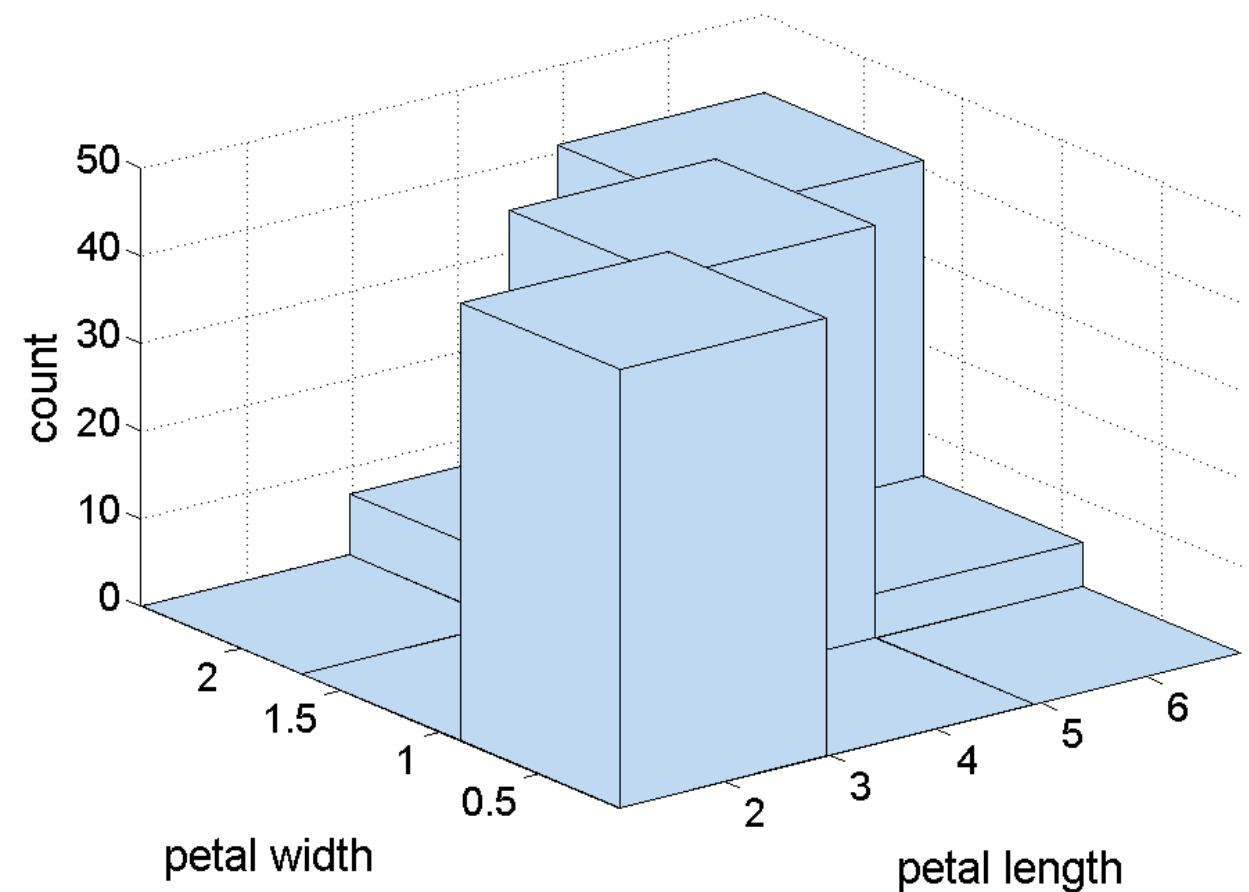
# Plot histogram

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

Two-Dimensional Histograms

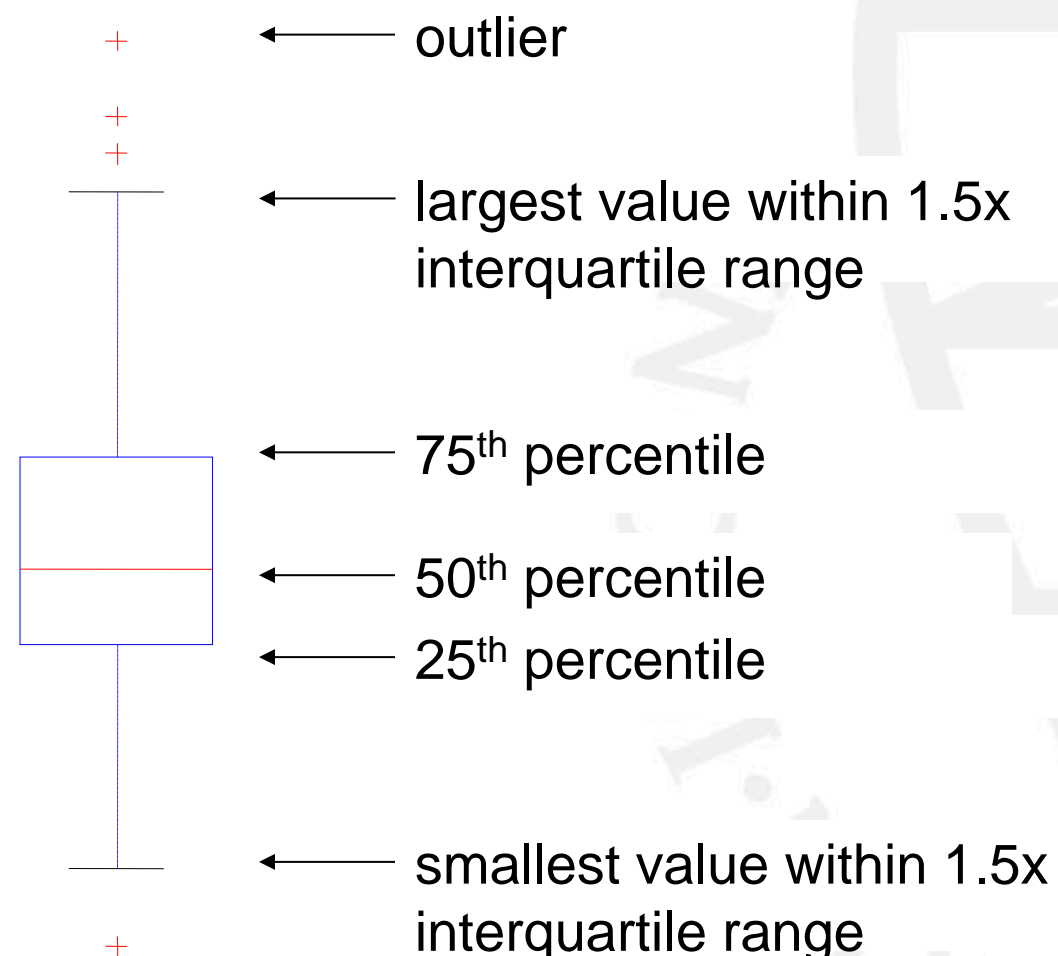
- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
- Code:

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



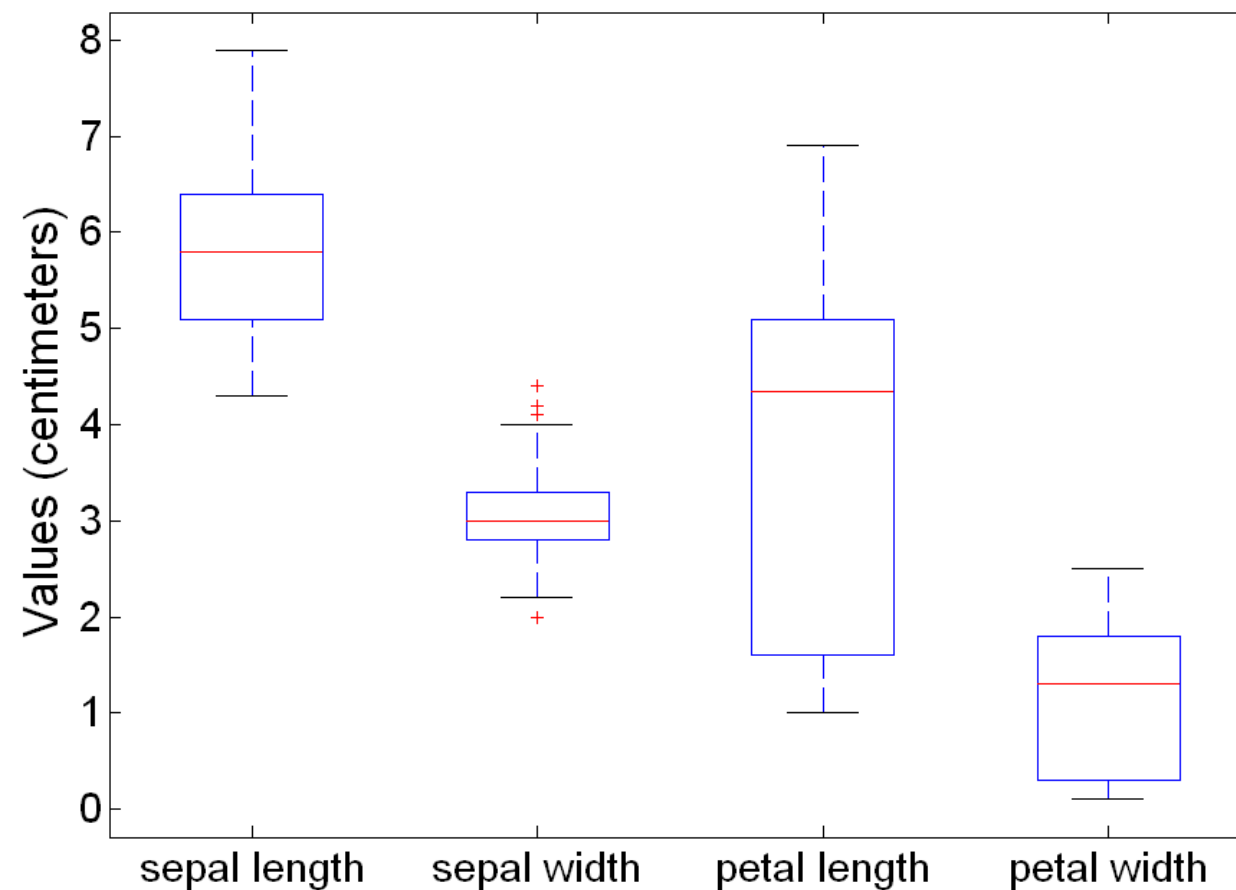
Visualization Techniques: Box Plots

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



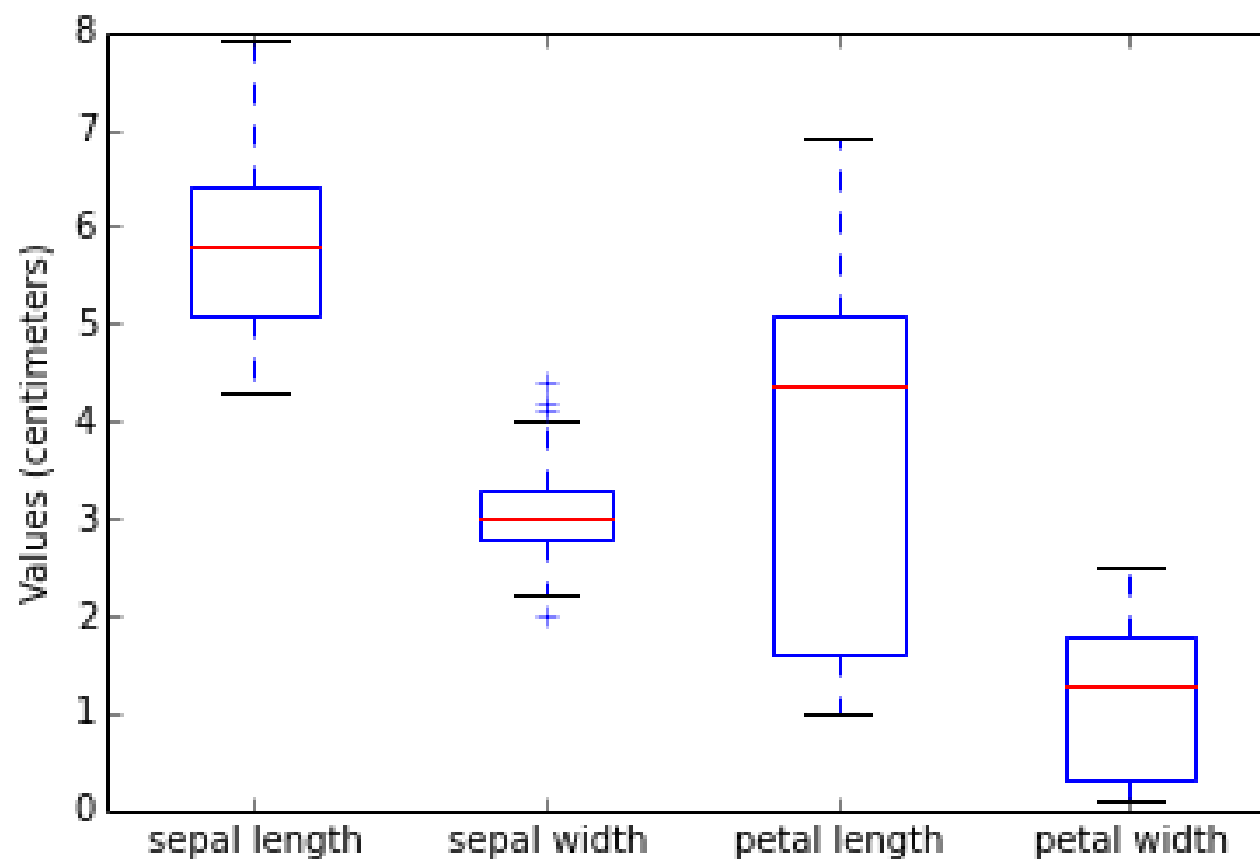
Example of Box Plots

```
varnames = ...  
    {'sepal length', 'sepal width', 'petal length', 'petal width'};  
boxplot(meas, varnames)  
ylabel('Values (centimeters)')
```

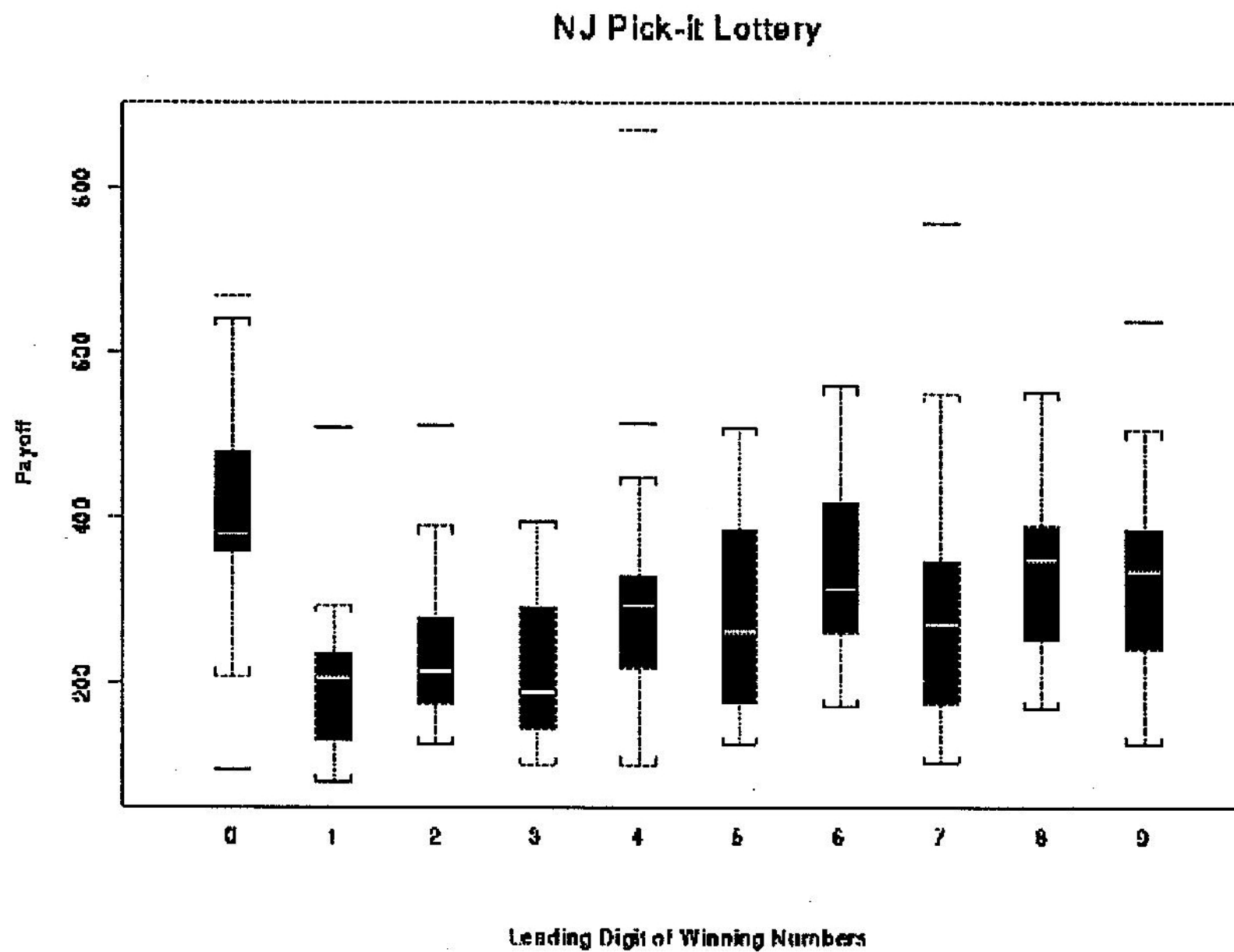


Box Plots in Python

```
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

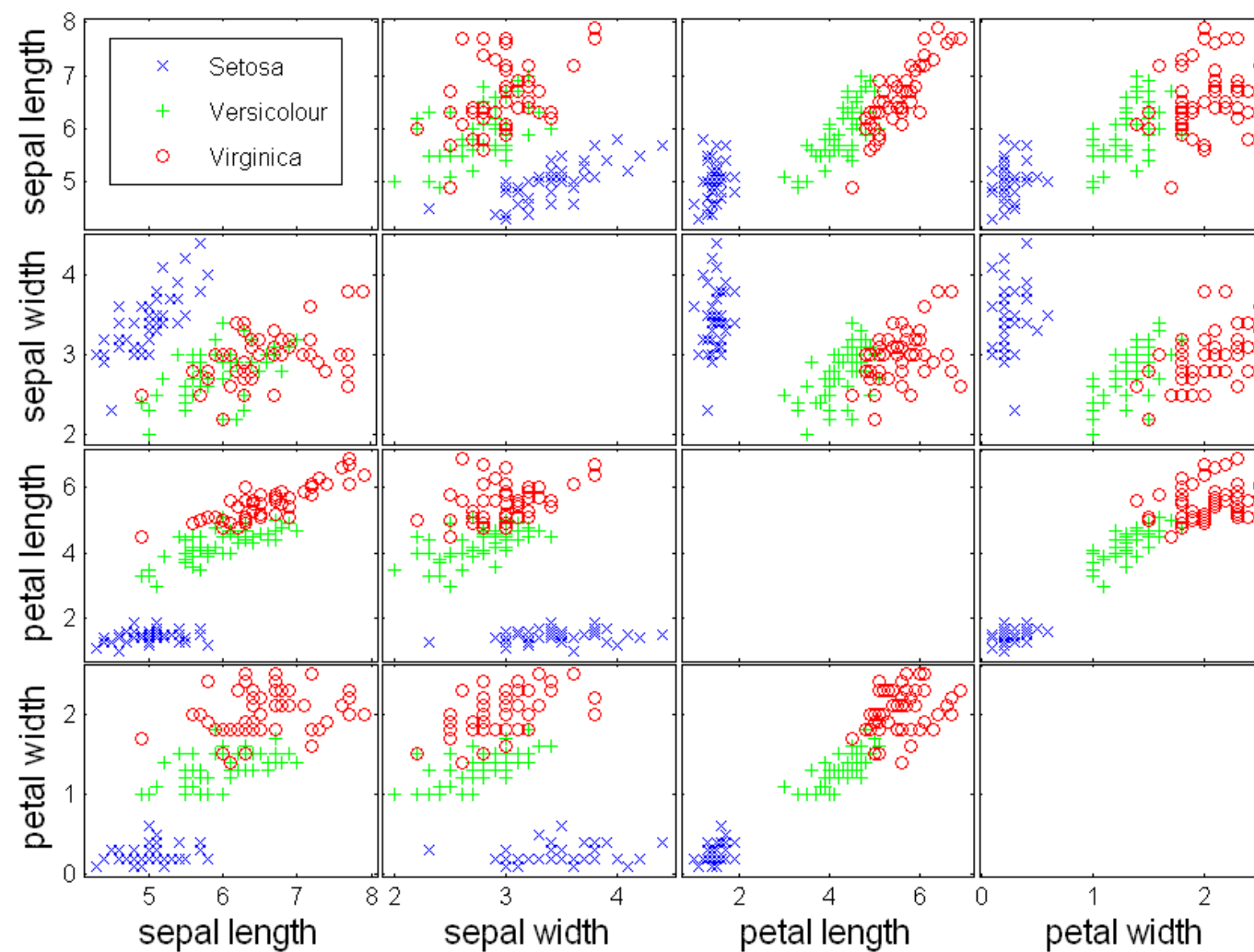


Visualization Techniques: Scatter Plots

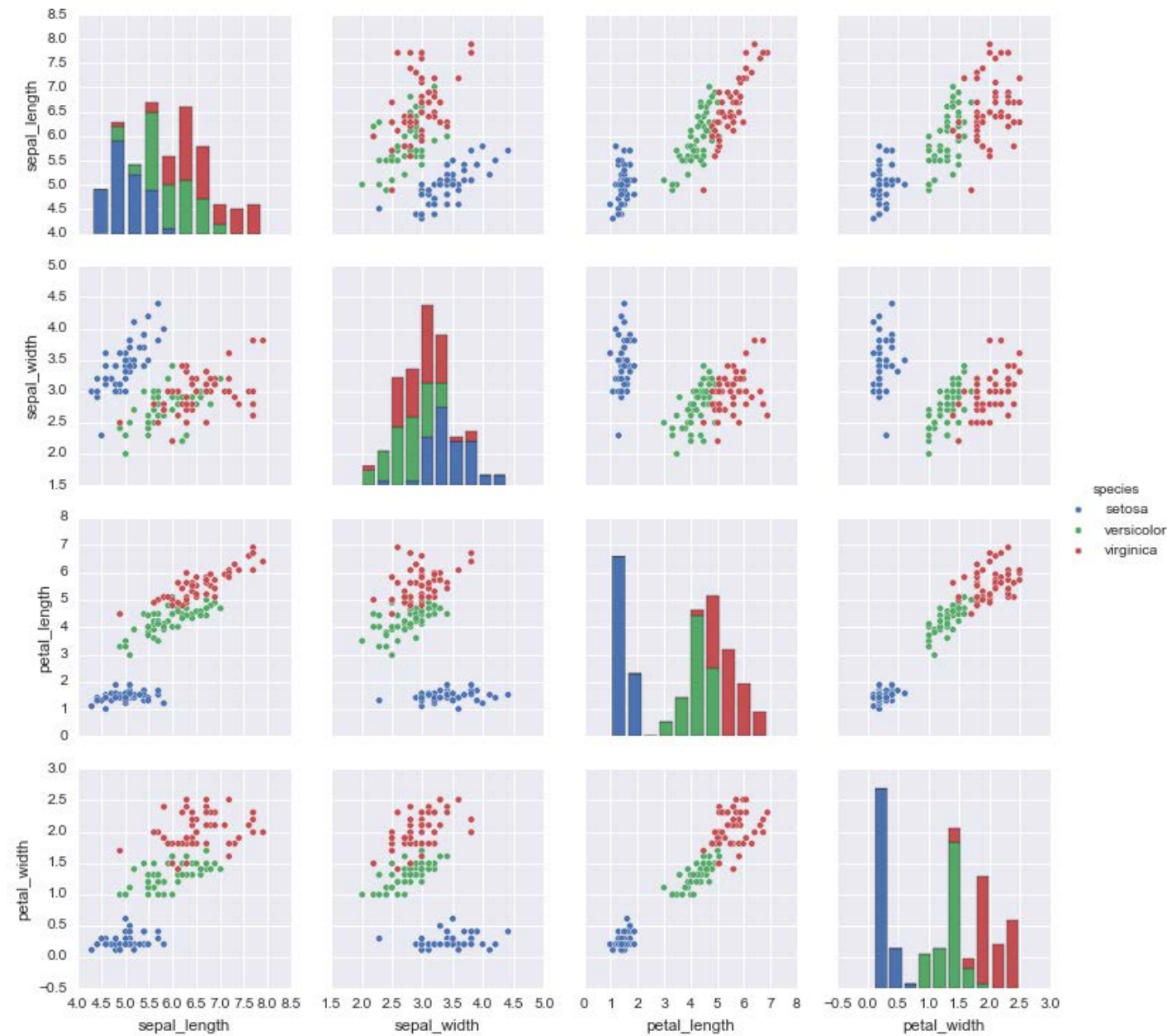
- Attributes values determine the position
- Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
- Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
- Arrays of scatter plots can compactly summarize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

Scatter Plot Array of Iris Attributes

```
gplotmatrix(meas,[],species,...  
            [],'x*o',[],'on','none',varnames)
```



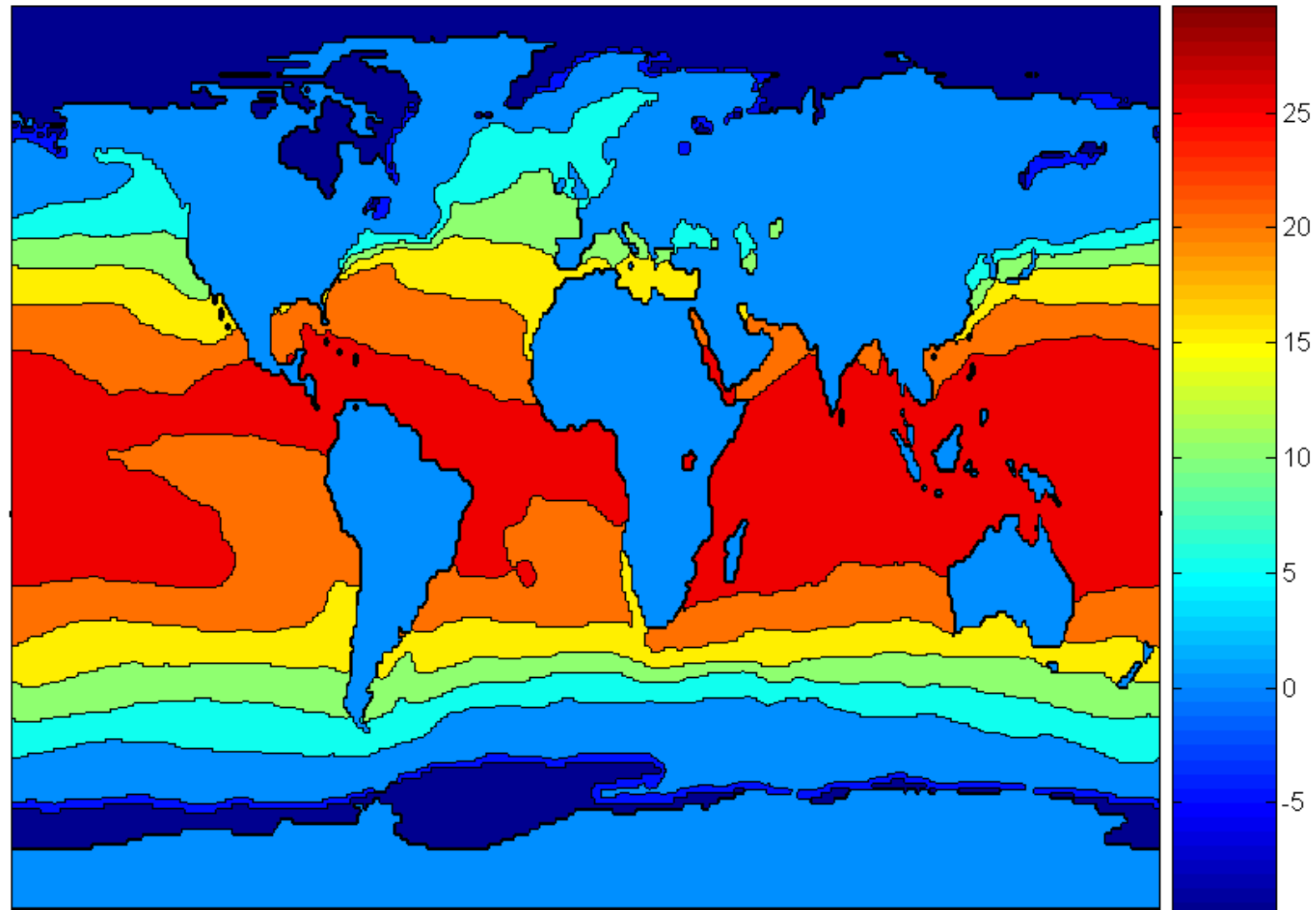
Python: use pandas or seaborn



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.

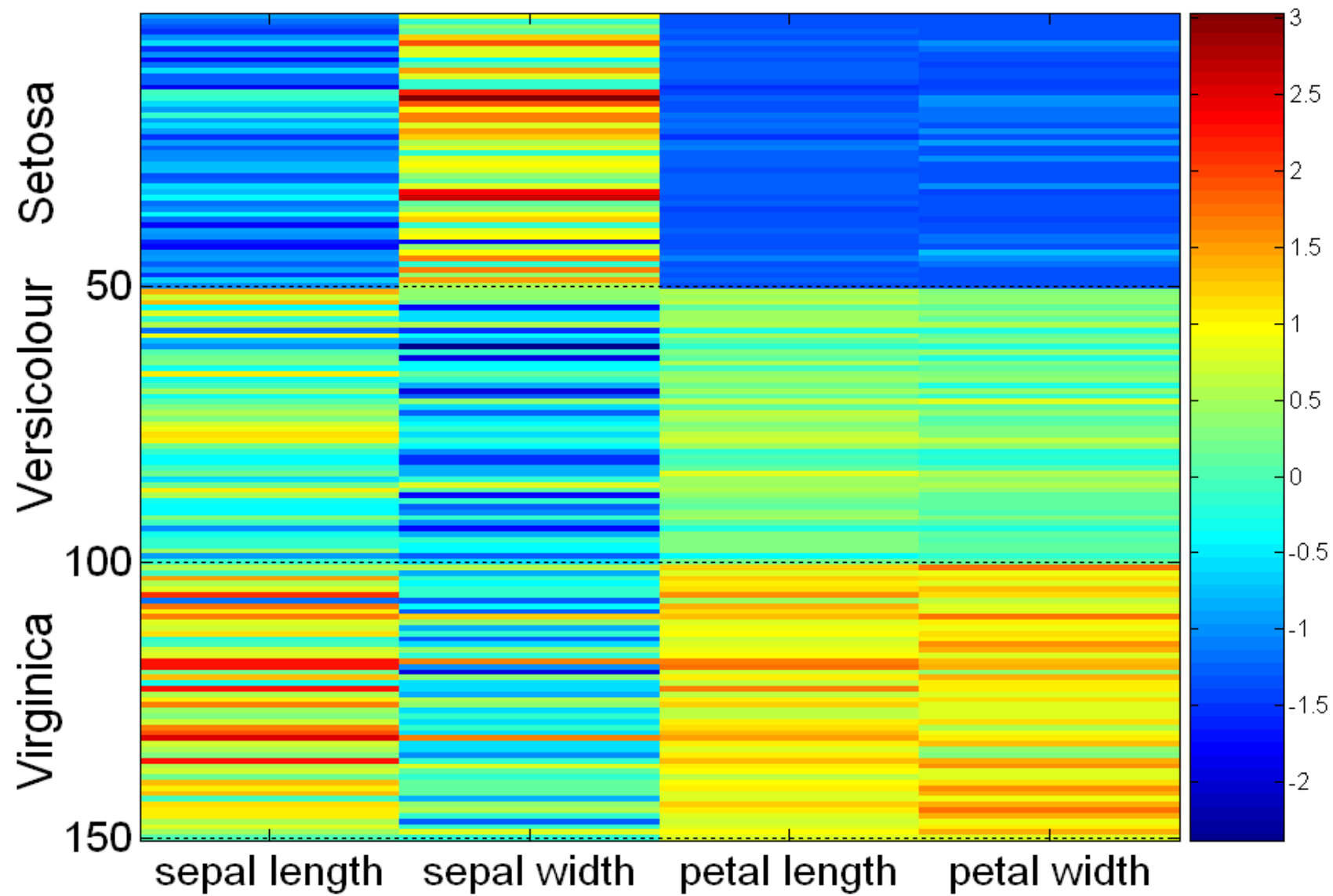
Contour Plot Example: SST Dec, 1998



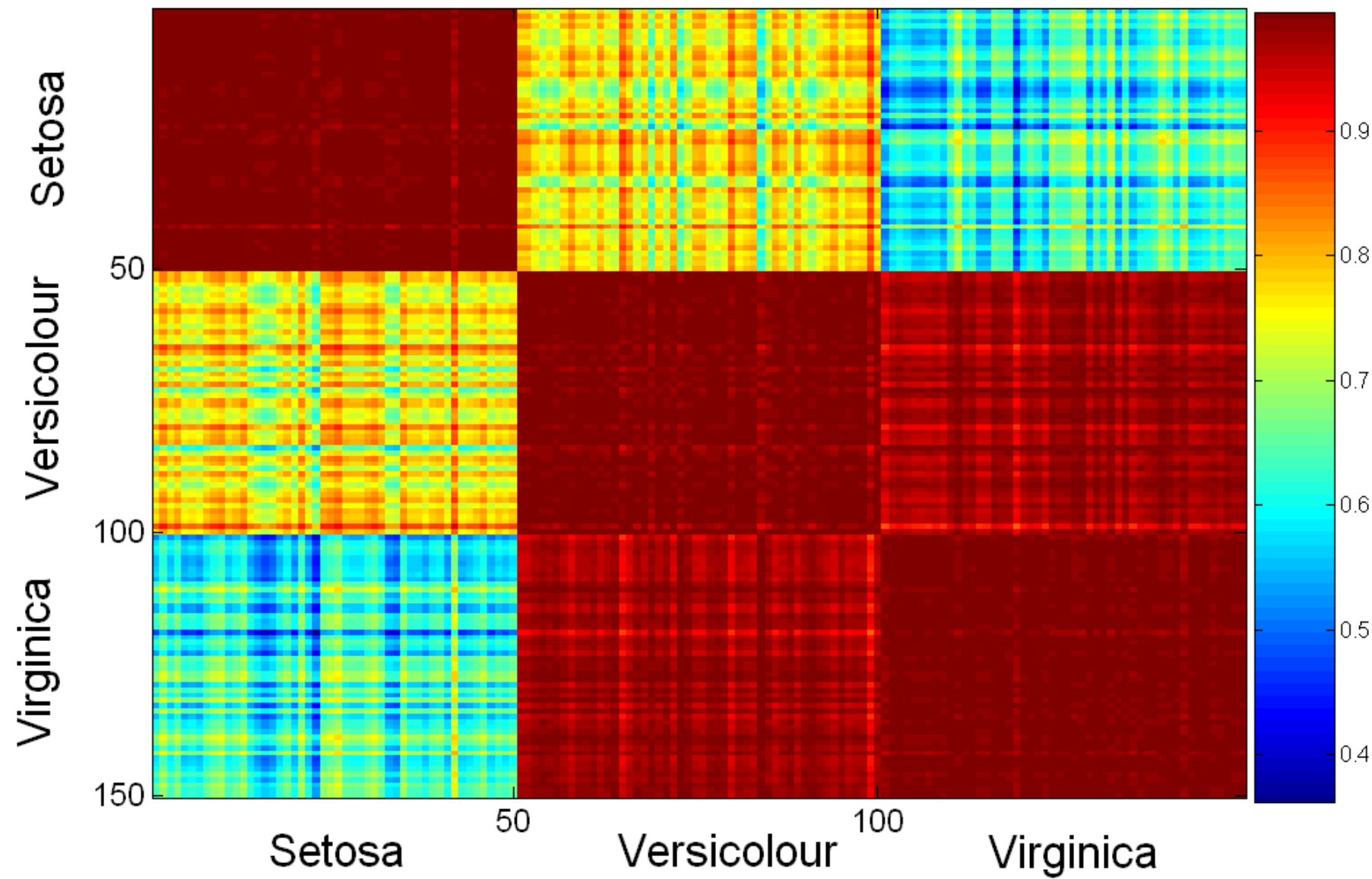
Visualization Techniques: Matrix Plots

- 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

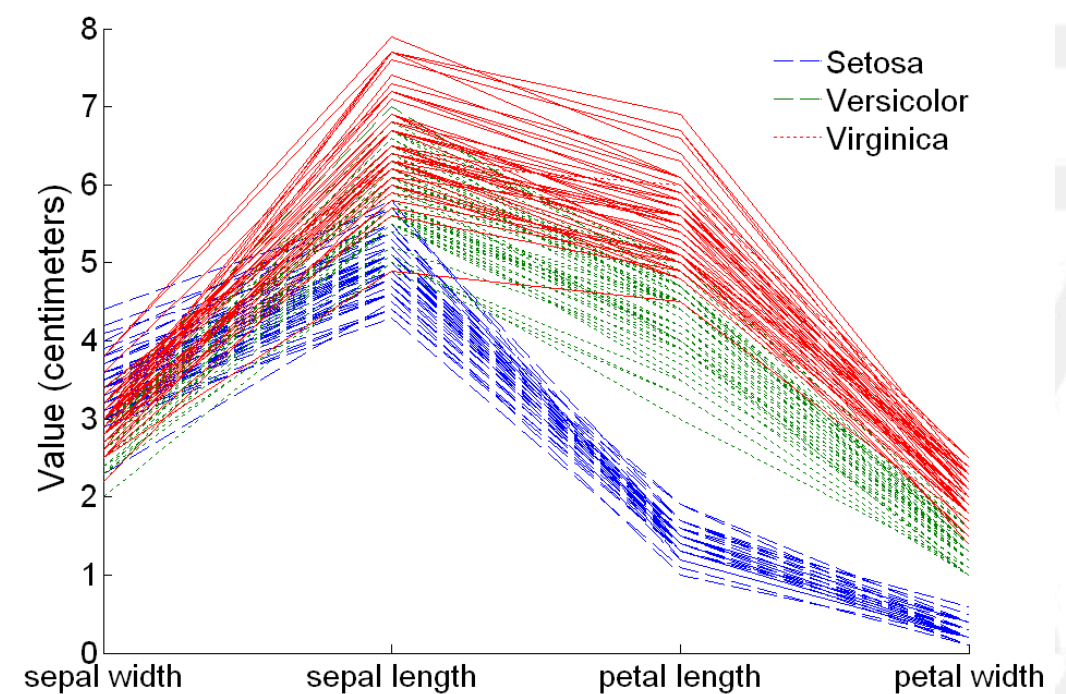
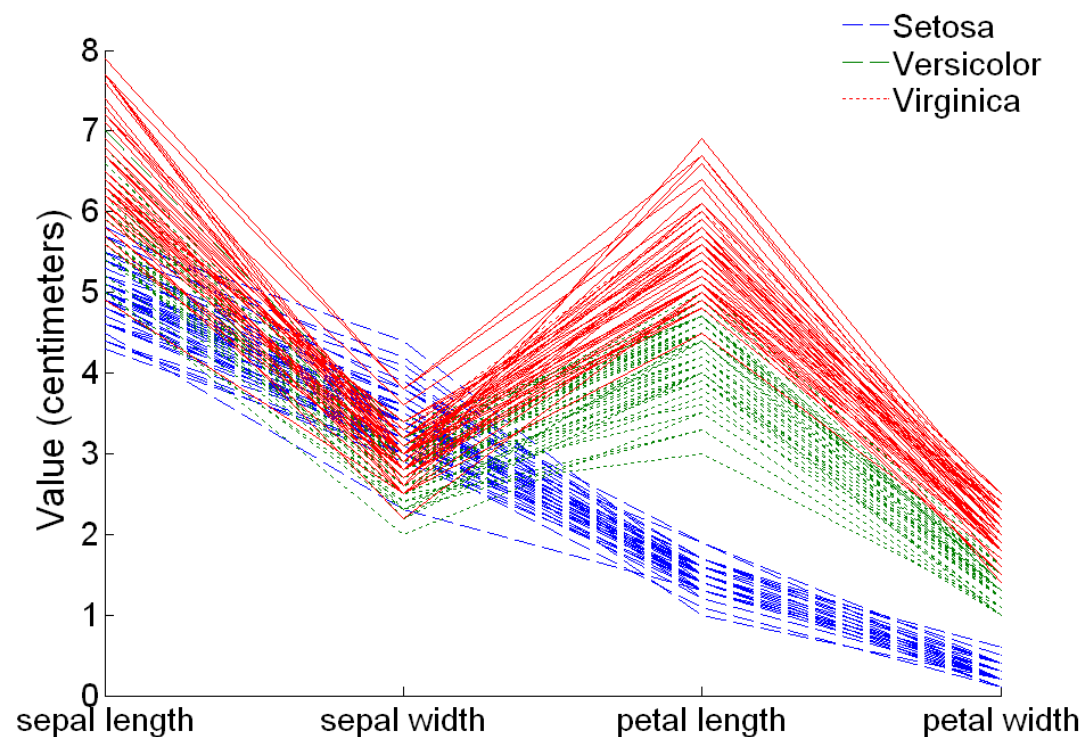


Visualization Techniques: Parallel Coordinates

- 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

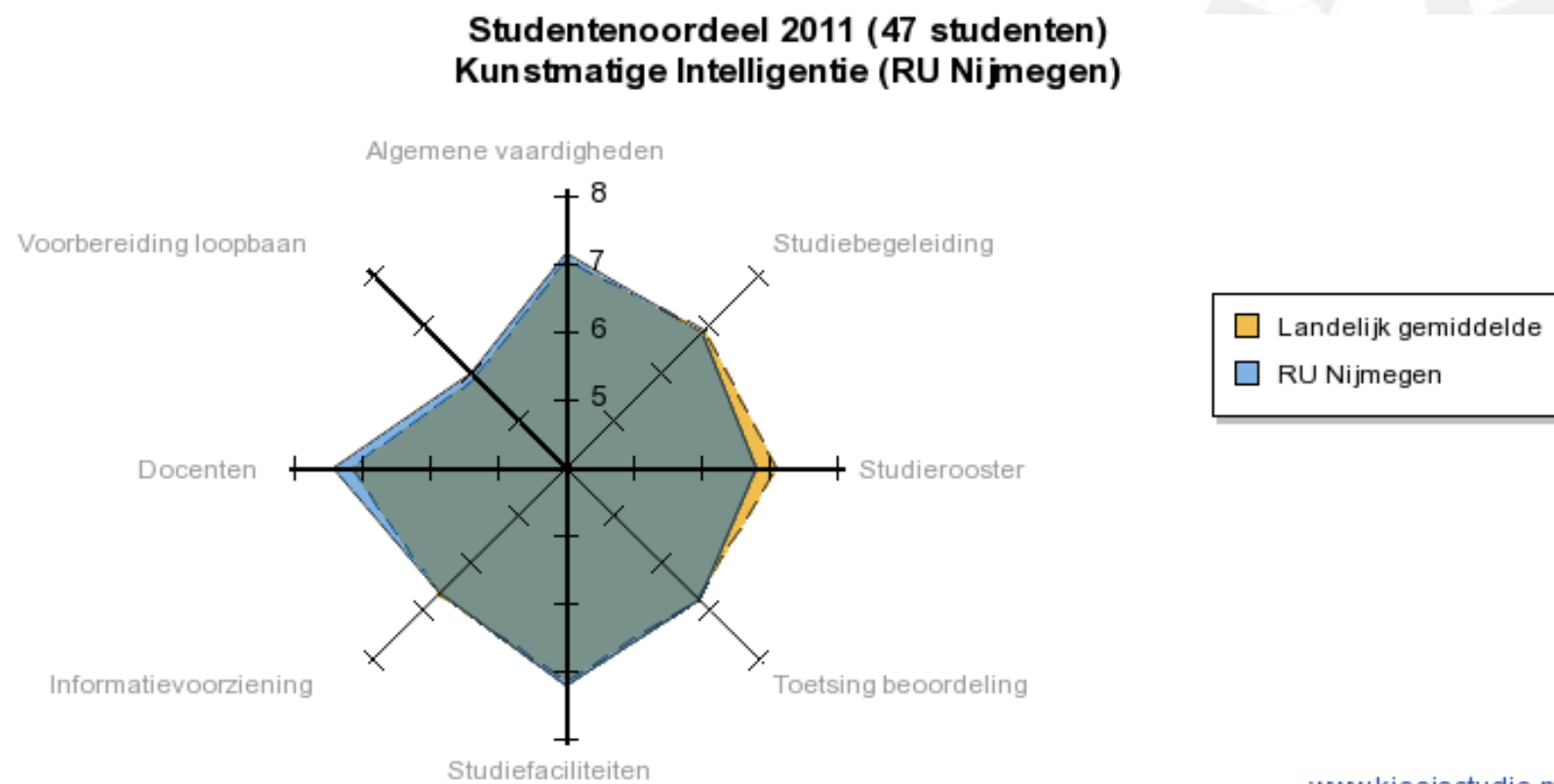
Parallel Coordinates Plots for Iris Data

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



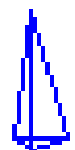
Star Plots

- 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)
```



1



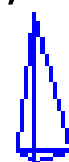
2



3

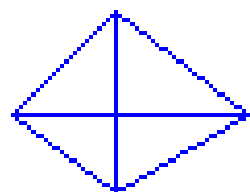


4

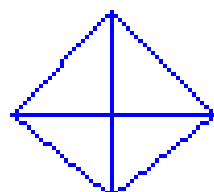


5

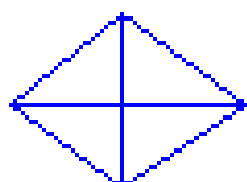
Setosa



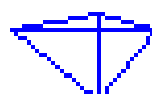
51



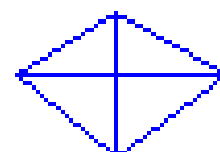
52



53

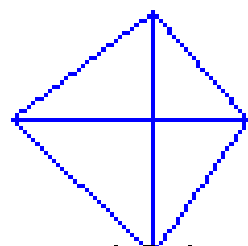


54

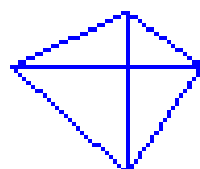


55

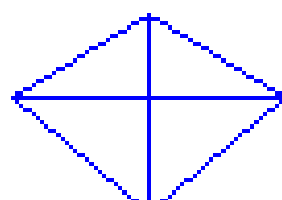
Versicolour



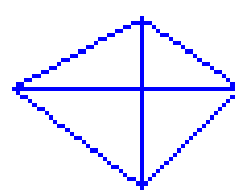
101



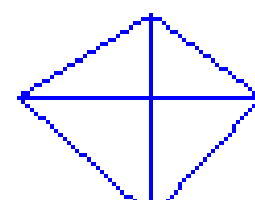
102



103



104



105

Virginica

MATLAB Code for Star Plot

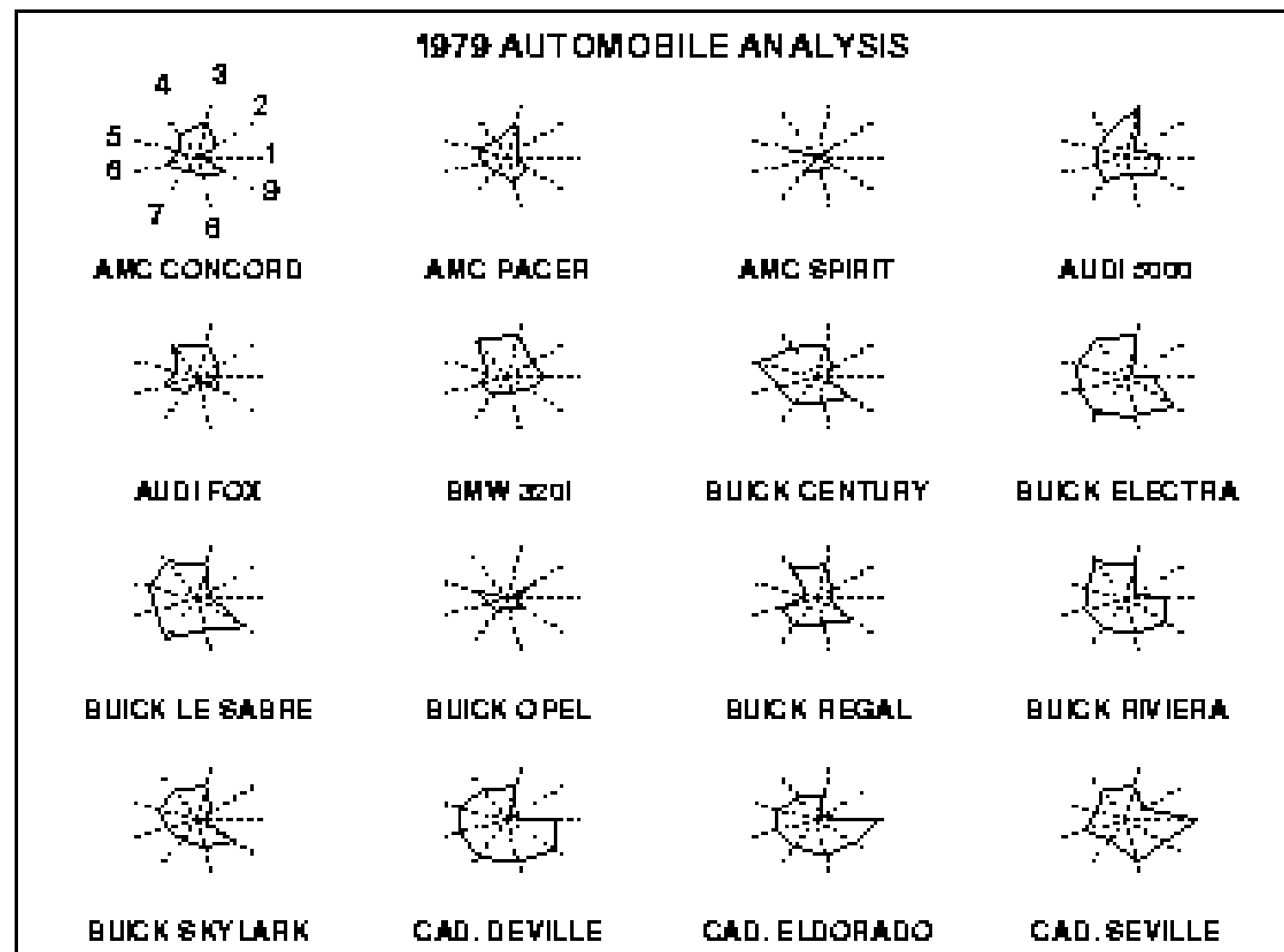
```
% Choose indices and create labels

indices = [1:5,51:55,101:105];
labels = cell(1,15);
for i=1:15,
    labels{i} = num2str(indices(i));
end

% Make star plots

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')
```



1



2



3

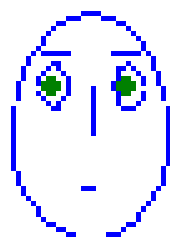


4



5

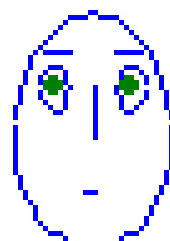
Setosa



51



52



53



54



55

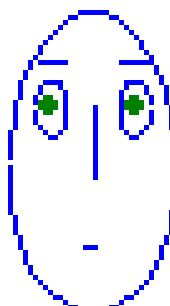
Versicolour



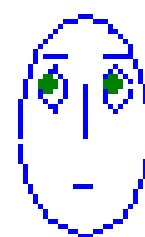
101



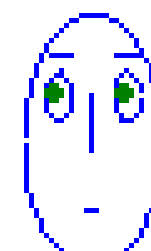
102



103



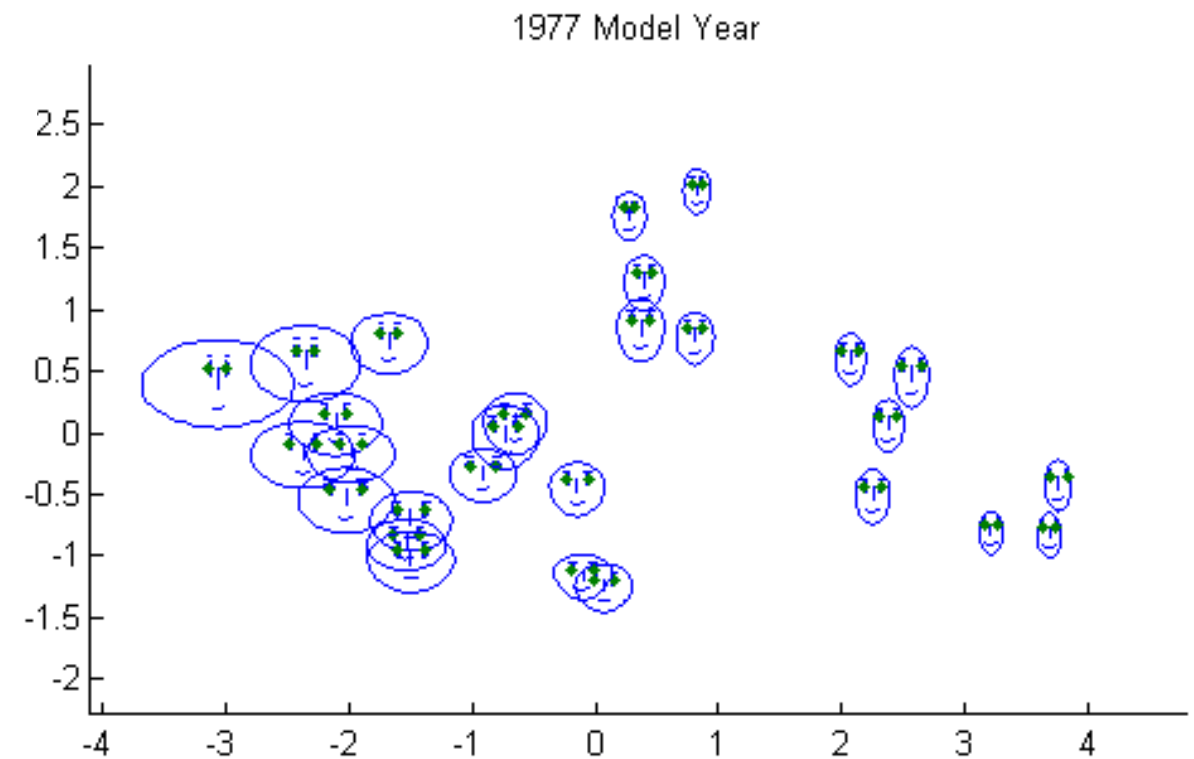
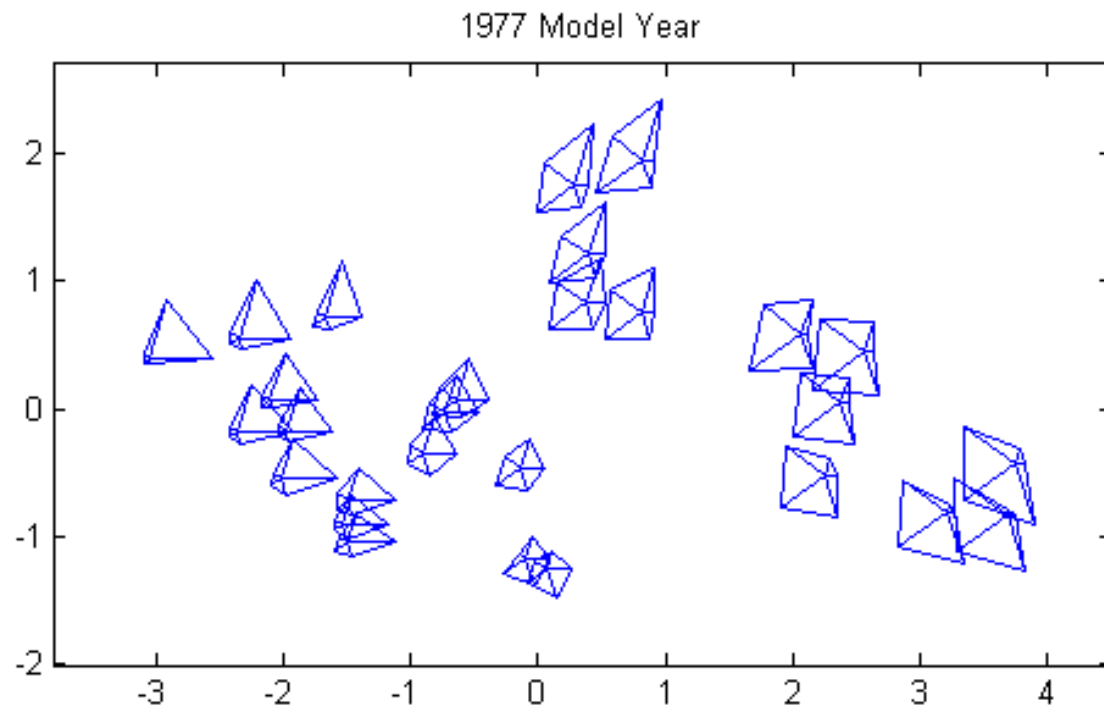
104



105

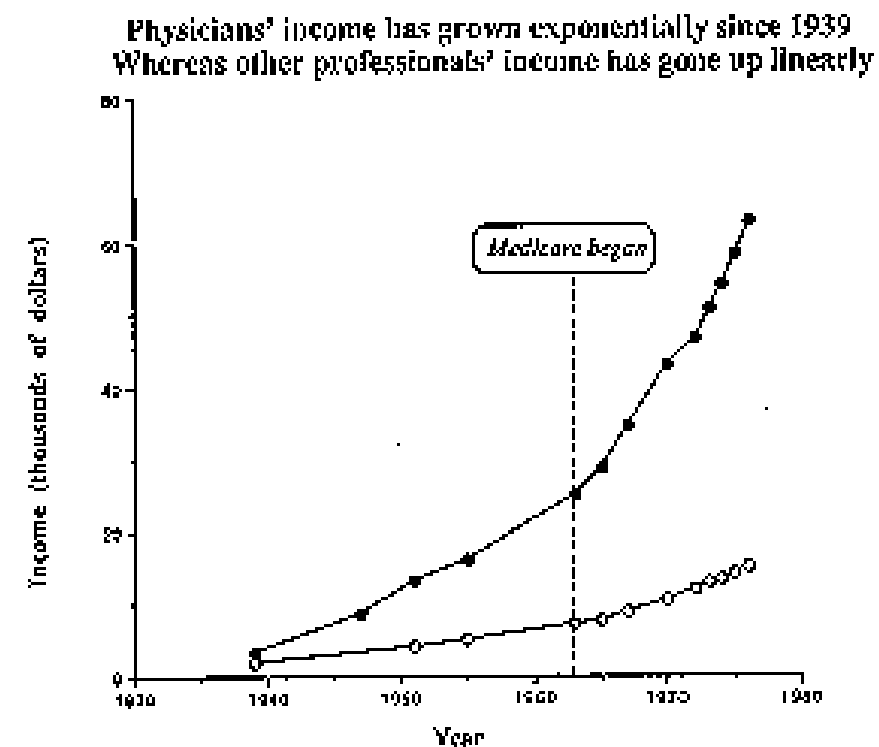
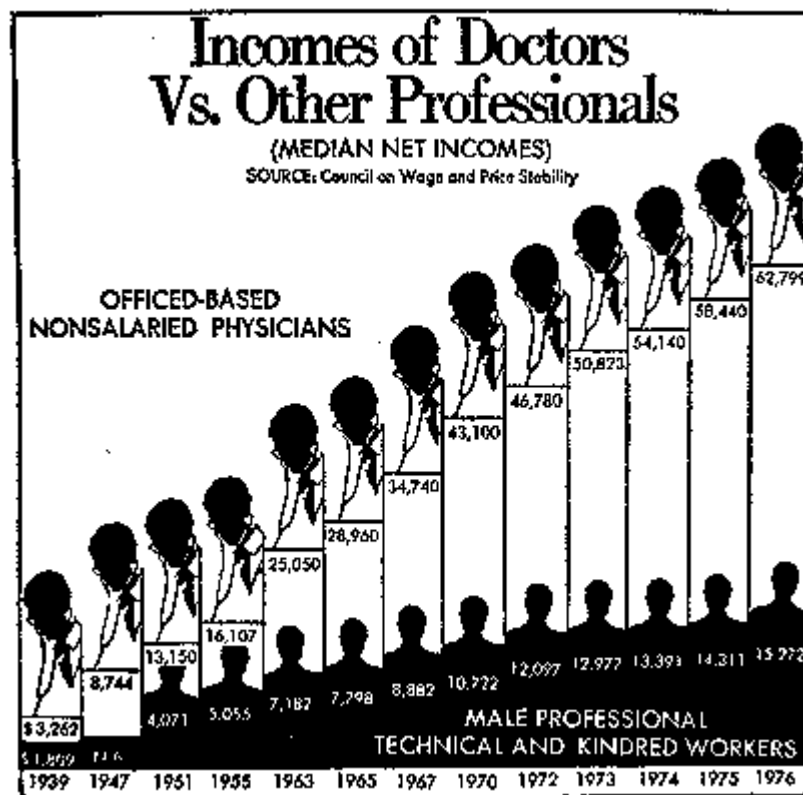
Virginica

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:

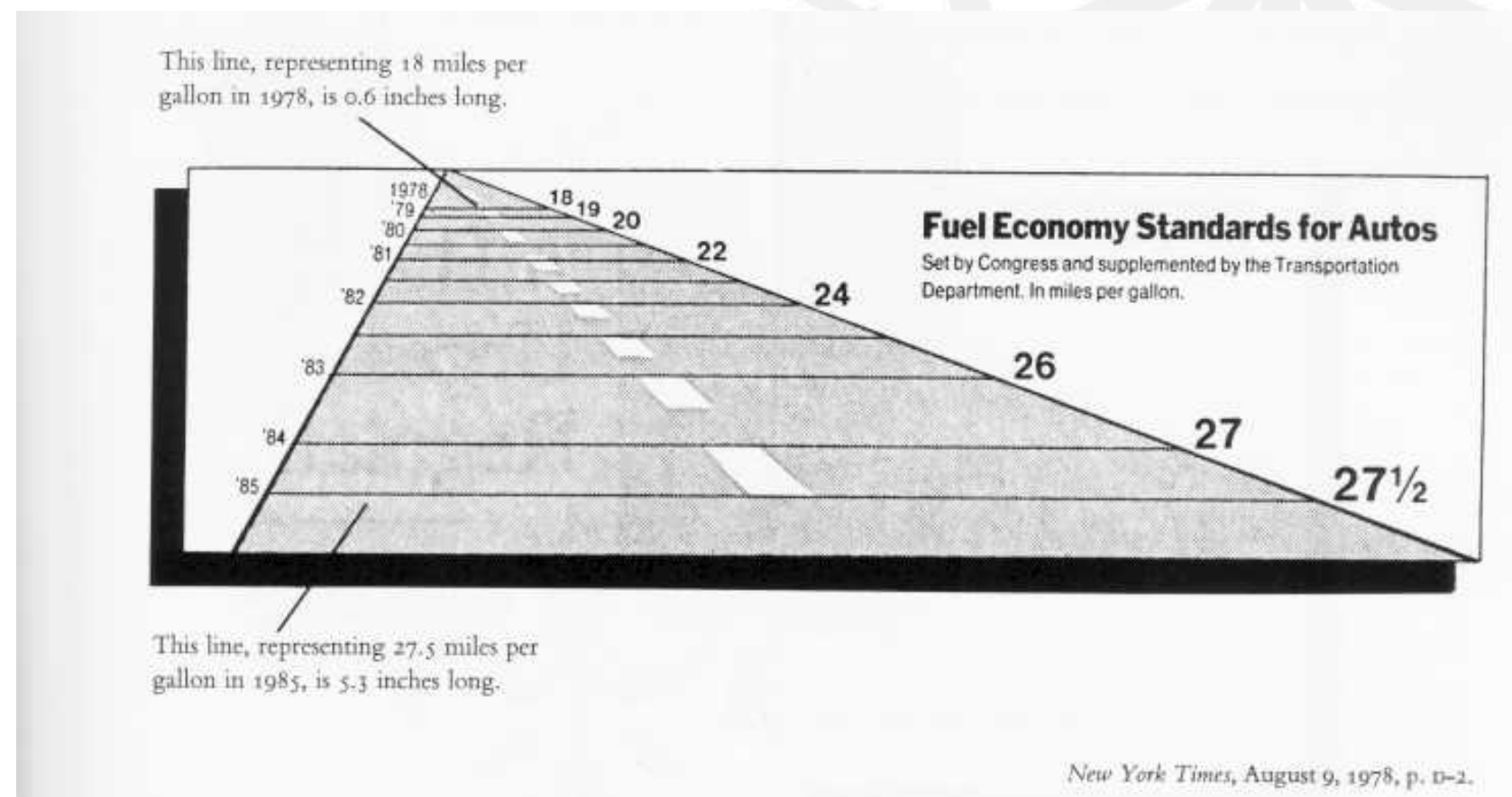
$$\text{Lie factor} = \frac{\text{size of effect shown in graph}}{\text{size of effect in data}}$$

- Where:

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

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.

THE SHRINKING FAMILY DOCTOR In California

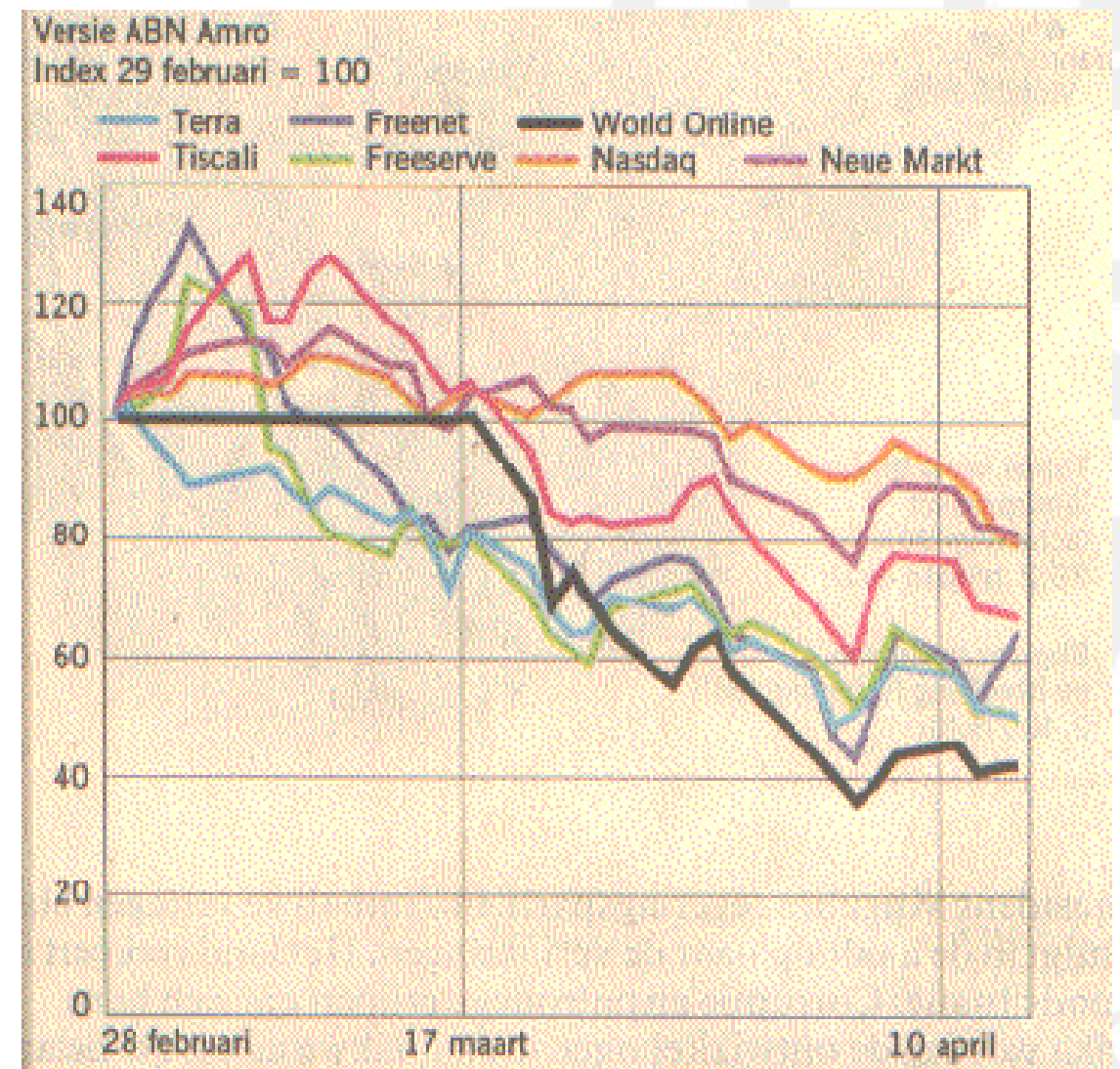
Percentage of Doctors Devoted Solely to Family Practice

1964	1975	1990
27%	16.0%	12.0%



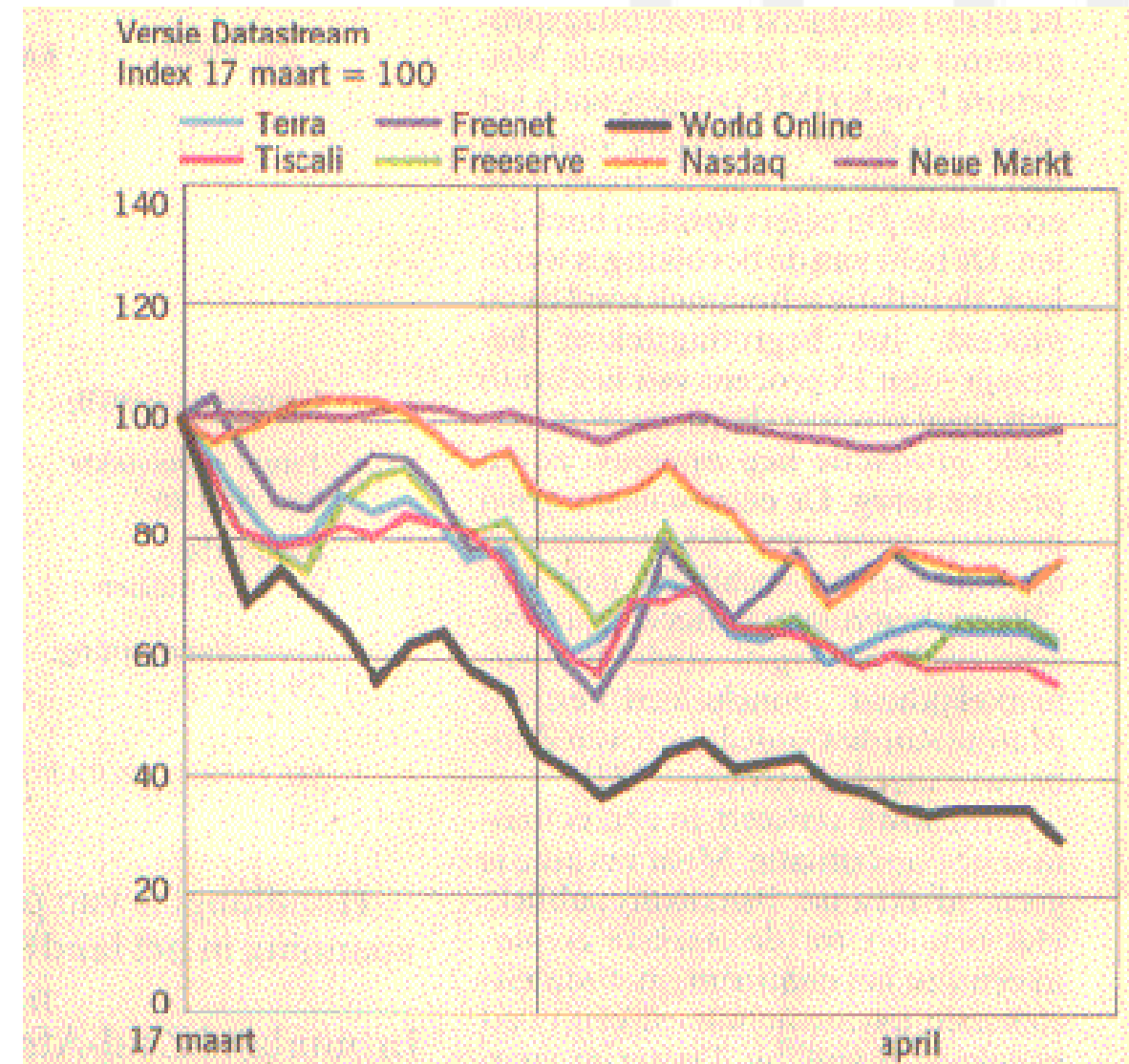
World On Line (1)

- 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)

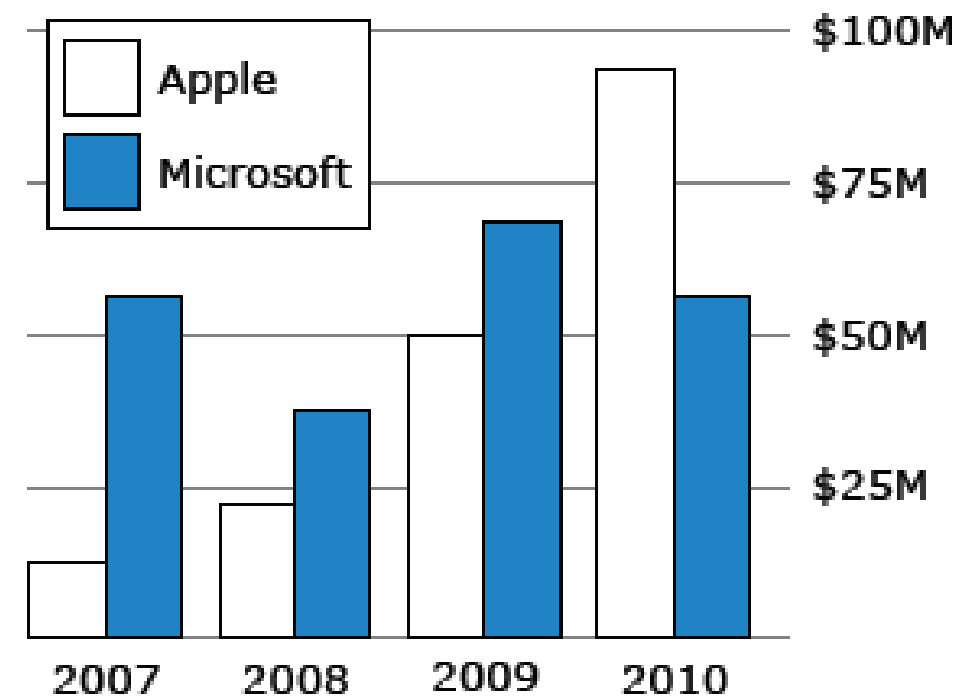
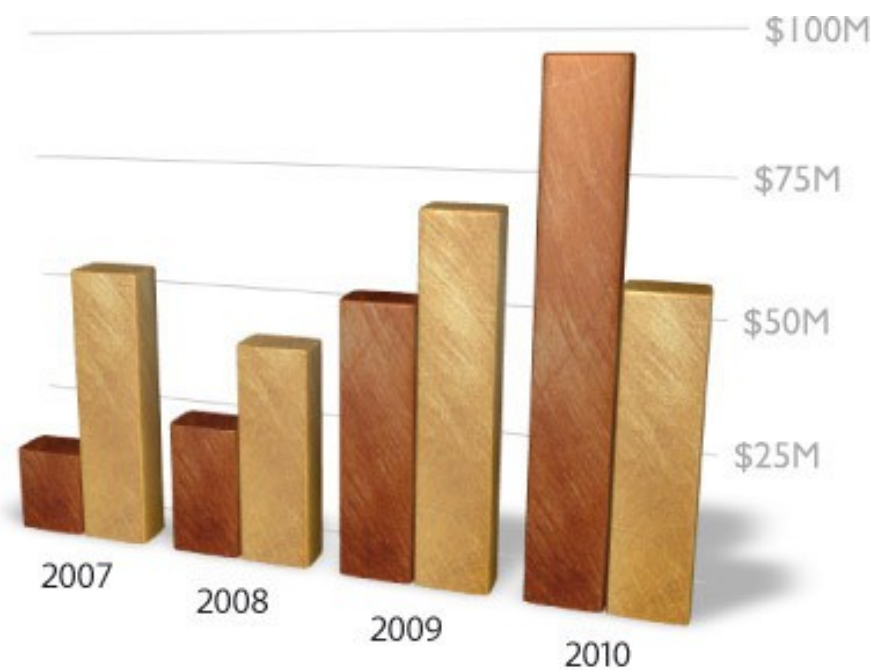
- No reason for a flat line between February 28 and March 17, the start of WOL's stock market quotation.
- Unshifting the base date for the index numbers to March 17 give this graph, in which WOL appears to be very quickly heading down the toilet...



ACCENT

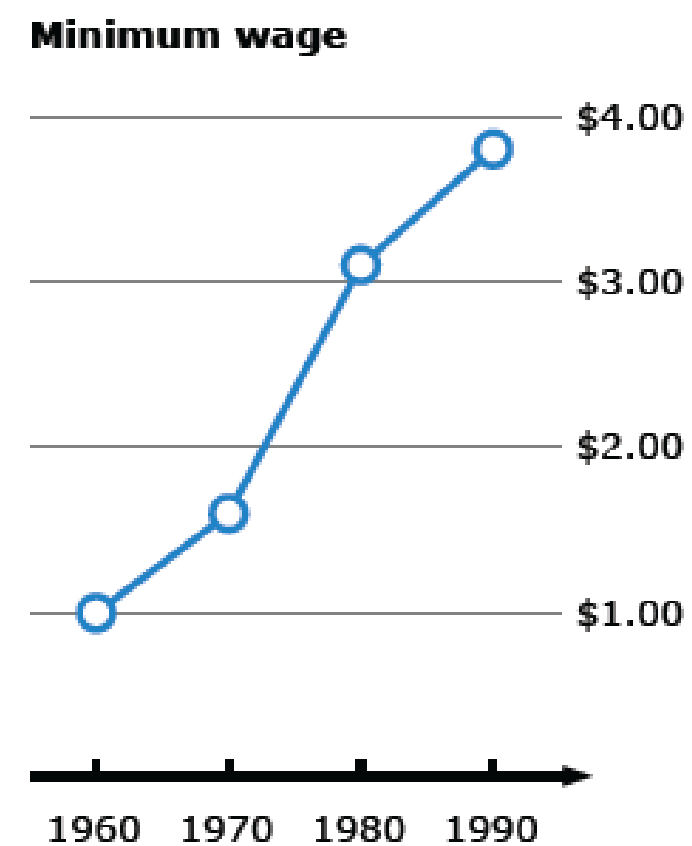
- Apprehension
 - Is it easy to see what is important in the graph?
- Clarity
 - Are the most important elements visually most prominent?
- Consistency
 - Have you used the same colors, shapes, etc. as in other graphs?
- Efficiency
 - Does it convey its information in the most simple and efficient way?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

Apple vs Microsoft

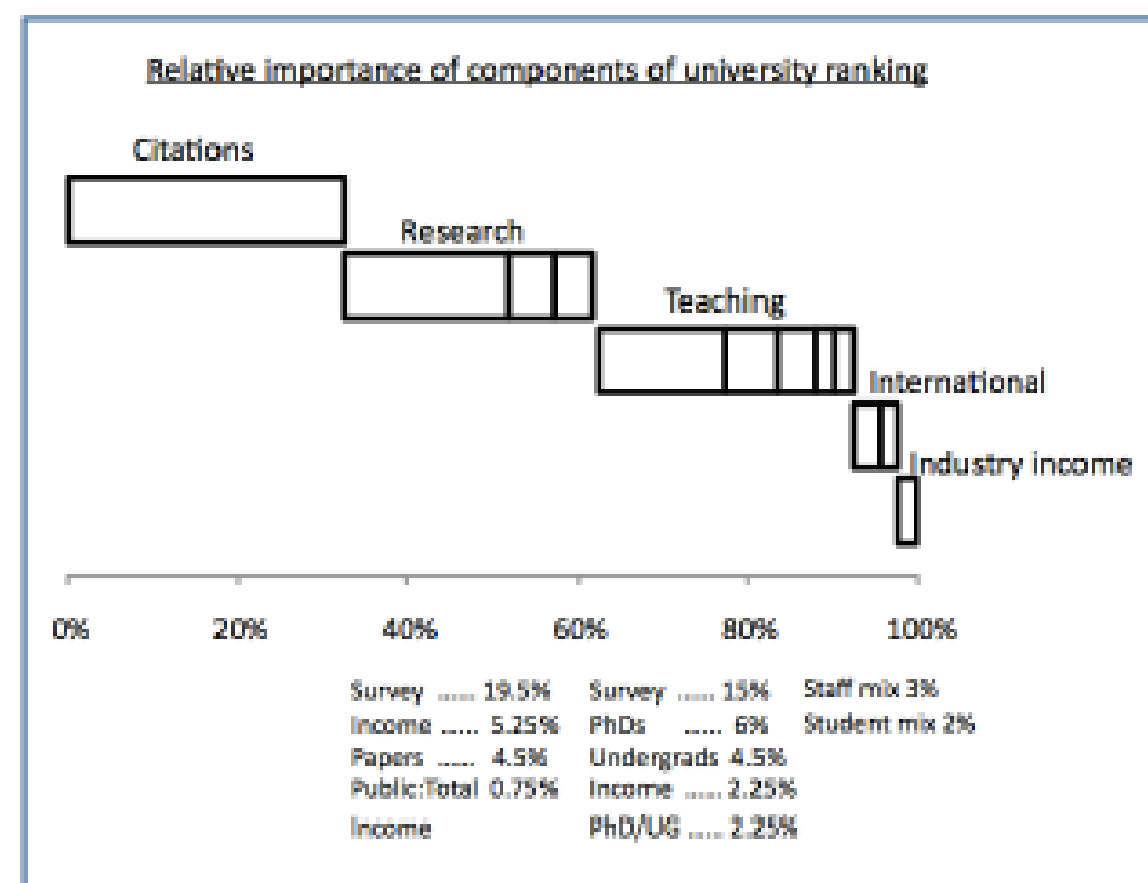
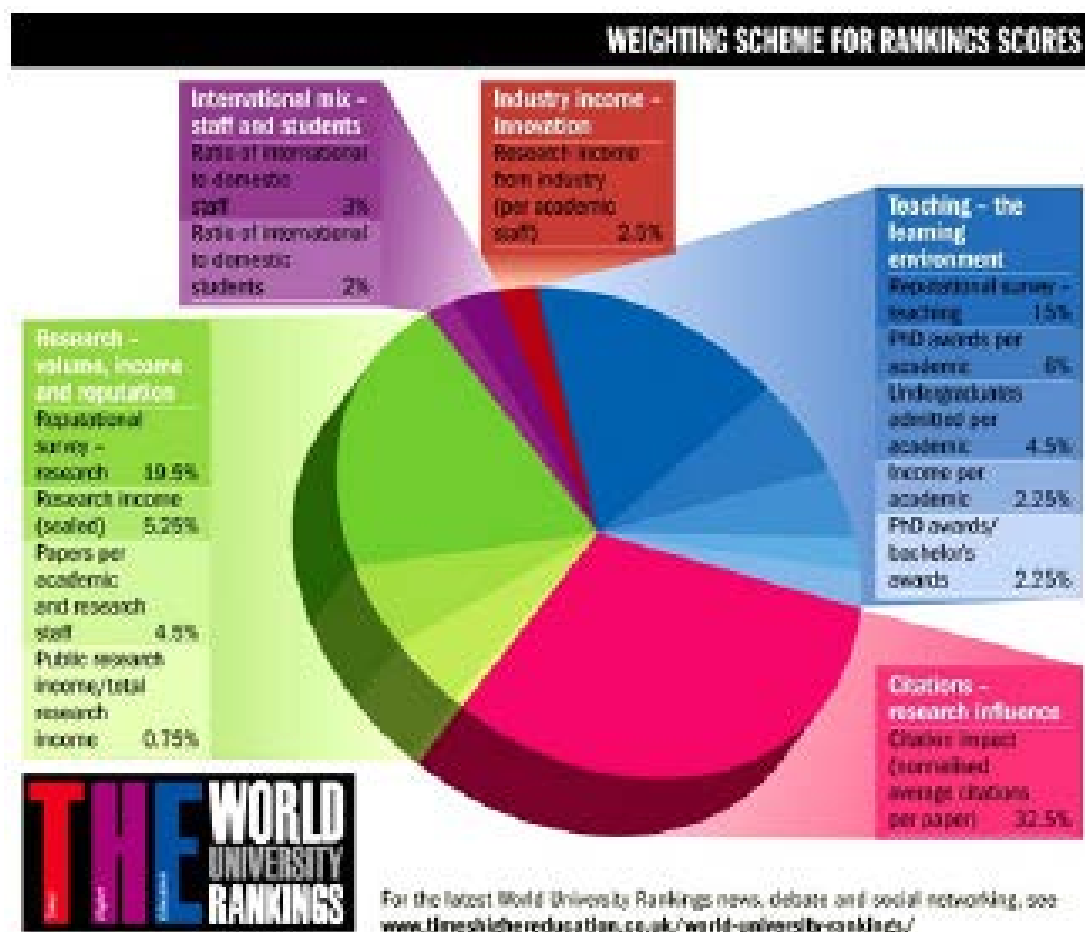


Minimum wage

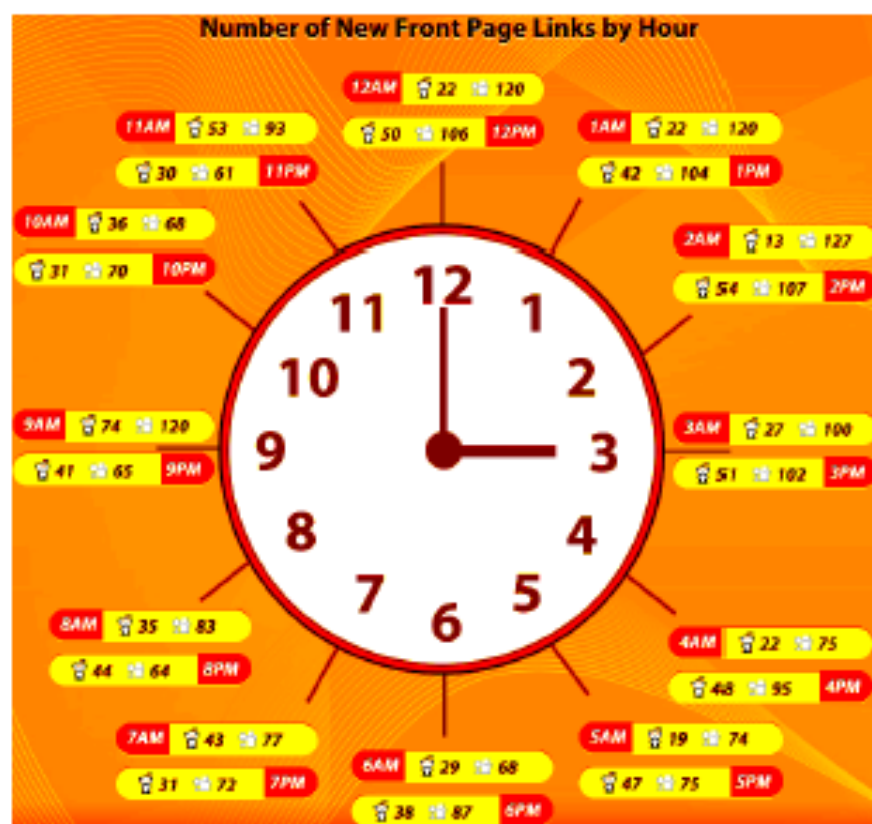
Minimum wage		
1960		\$1.00
1970		\$1.60
1980		\$3.10
1990		\$3.80



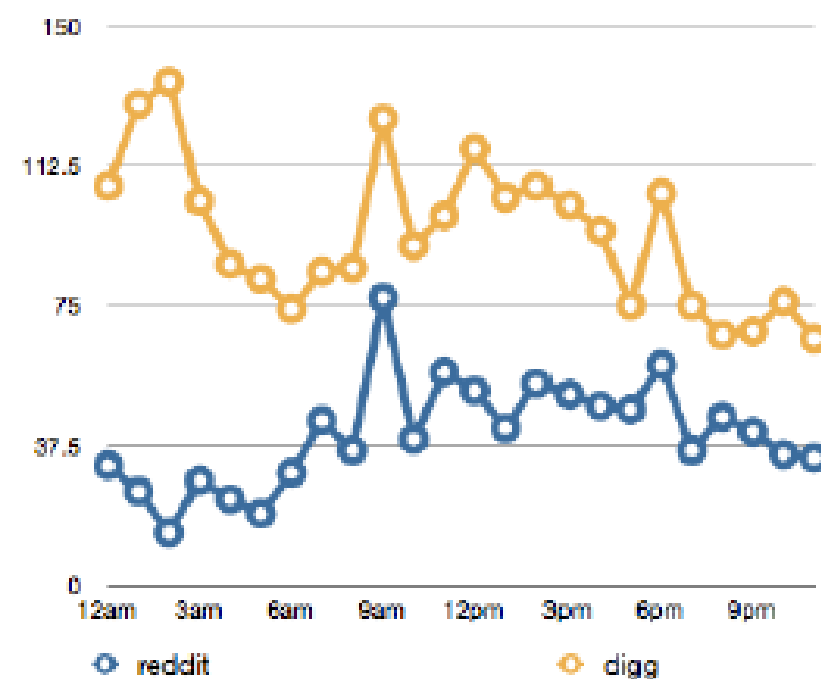
University Rankings



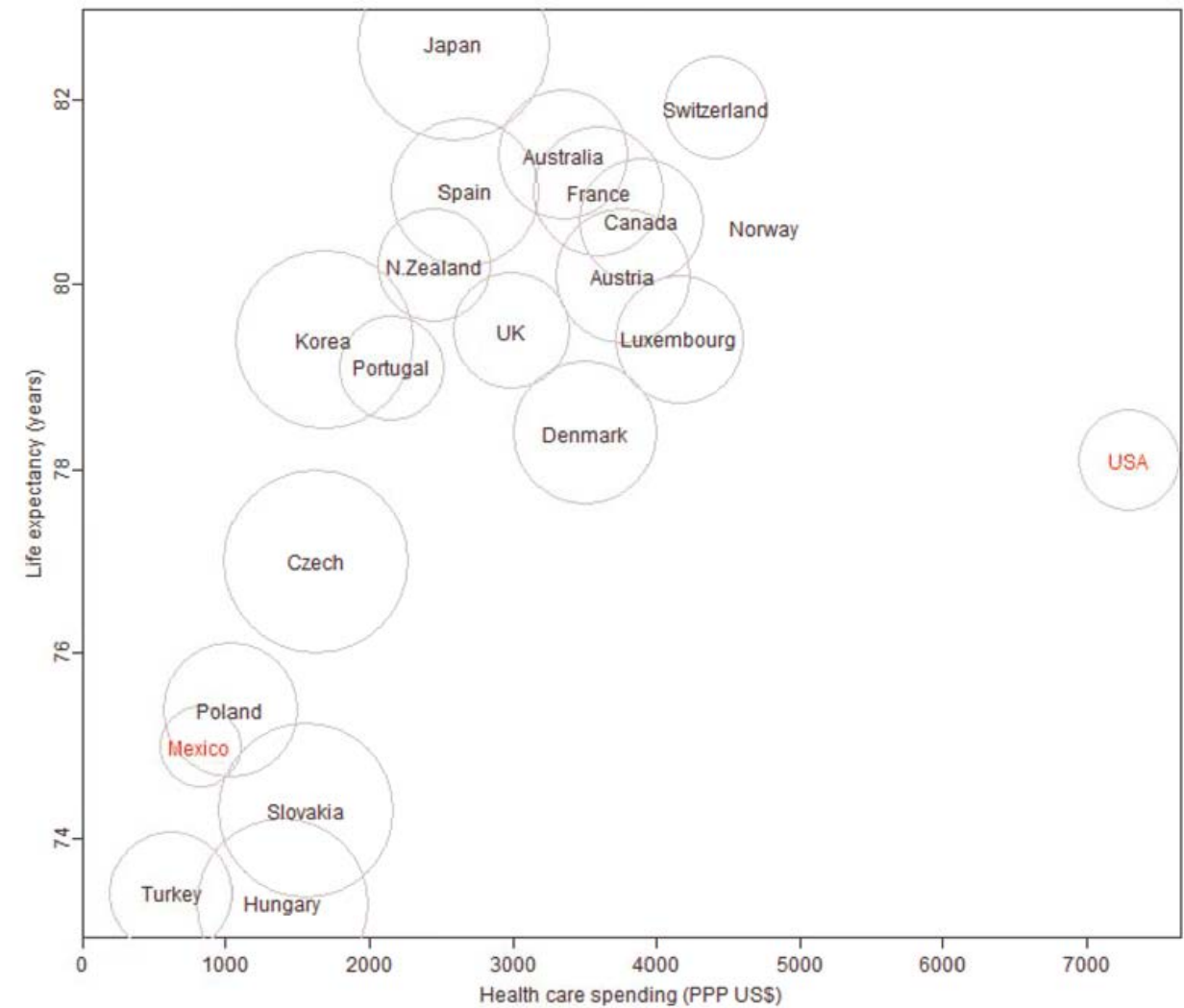
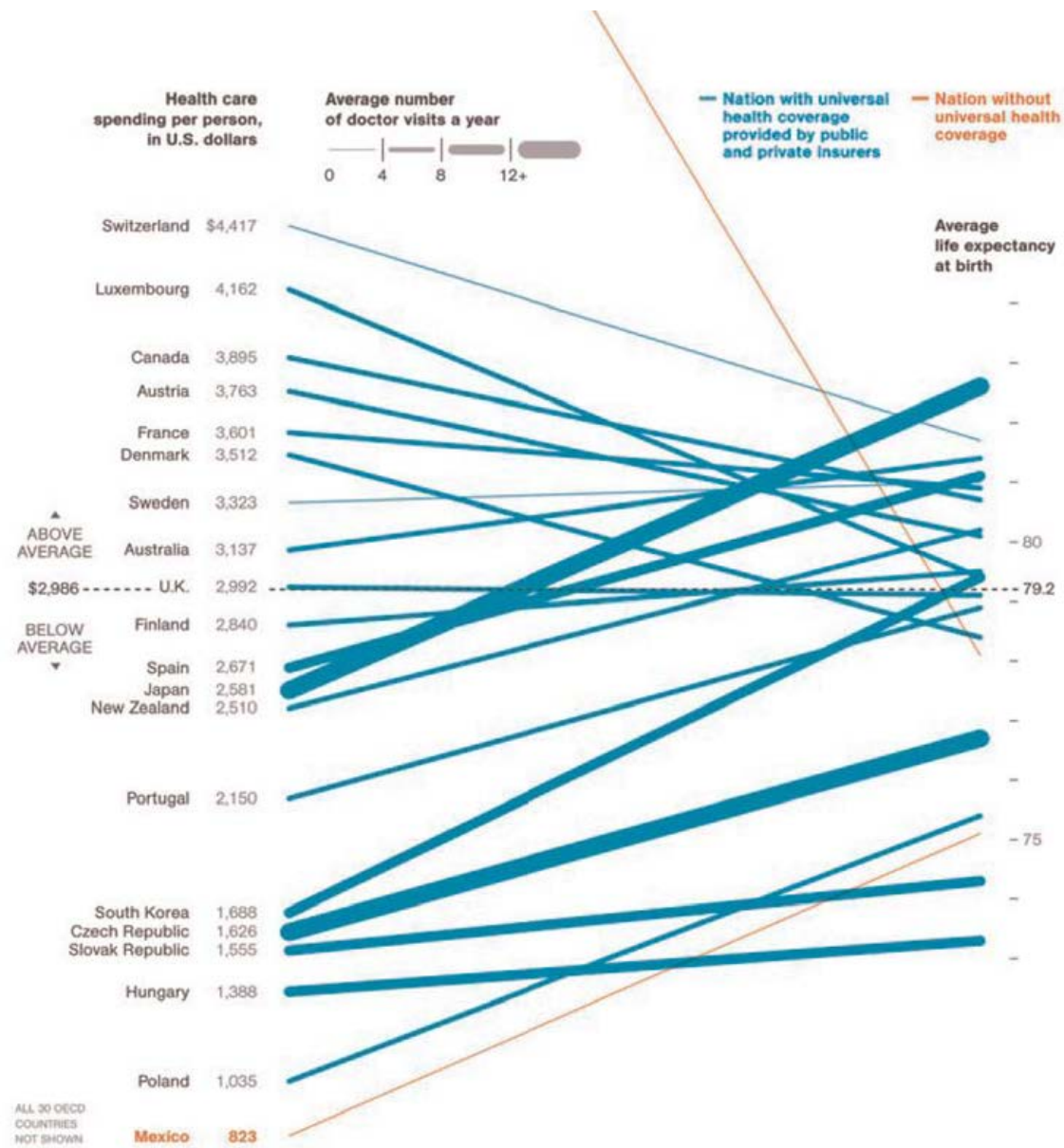
New Links



Number of New Front Page Links by Hour

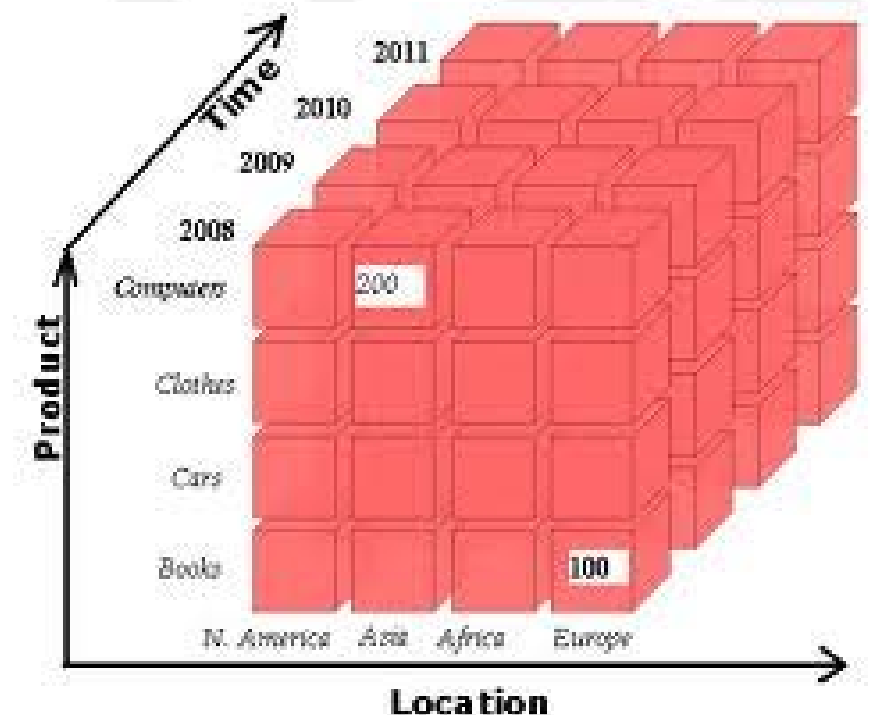


Life Expectancy



OLAP

- On-Line Analytical Processing (OLAP) was proposed by E. F. Codd, the father of the relational database
- Relational databases put data into tables, while OLAP uses a multidimensional array representation.
 - Such representations of data previously existed in statistics and other fields
- There are a number of data analysis and data exploration operations that are easier with such a data representation:
 - dicing
 - slicing
 - drill down
 - roll up



Handling Big Data

- Data everywhere
 - Global data volume grows exponentially
 - Need means of economically storing and processing large data sets
- Opportunity
 - Commodity hardware is ultra cheap
 - CPU and storage even cheaper
- Traditional solution
 - Store data in a (relational) database
 - Run batch jobs for processing
- Problems with existing solutions
 - Databases are seek heavy; seeks are wasted time
 - Databases do not play well with commoditized hardware
 - Databases were not built with horizontal scaling in mind

Slides acknowledgment:
Friso van Vollenhoven, Xebia

Parallel Processing

- Eliminate the seeks, only sequential reading / writing
- Work with batches for efficiency
- Parallelize work load
- Distribute processing and storage

MapReduce and Hadoop

- 2000: Apache Lucene: batch index updates and sort/merge with on disk index
- 2002: Apache Nutch: distributed, scalable open source web crawler; sort/merge optimization applies
- 2004: Google publishes Google File System (GFS) and MapReduce (MR) papers
- 2006: Apache Hadoop: open source Java implementation of GFS and MR to solve Nutch' problem; later becomes standalone project

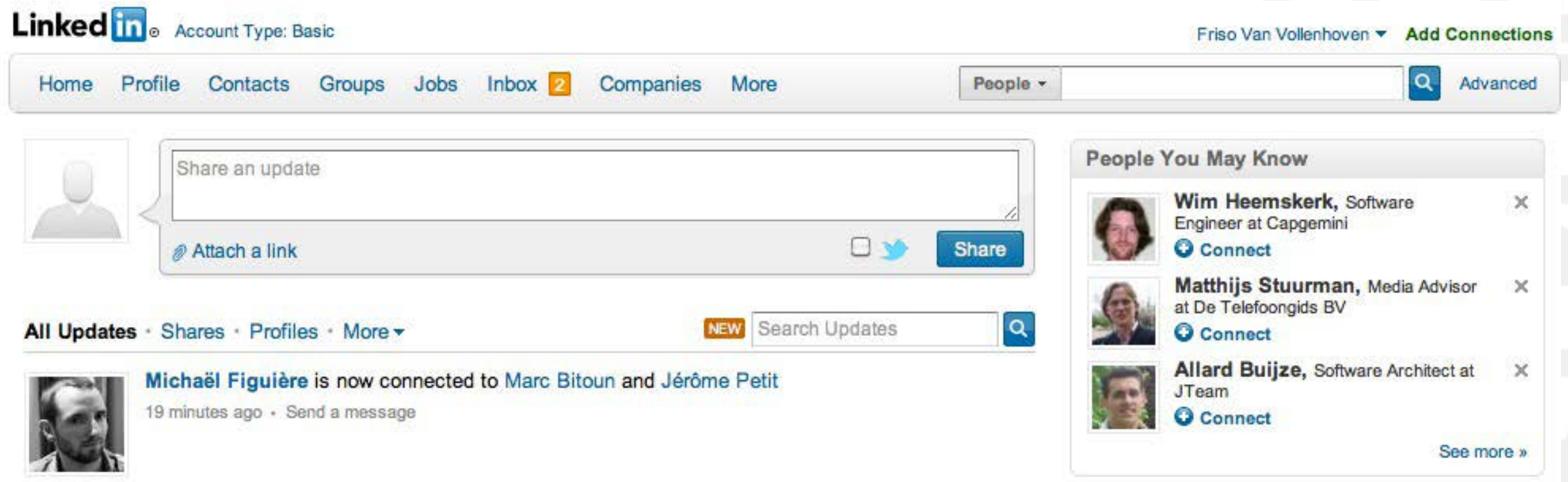


Hadoop Example (1)



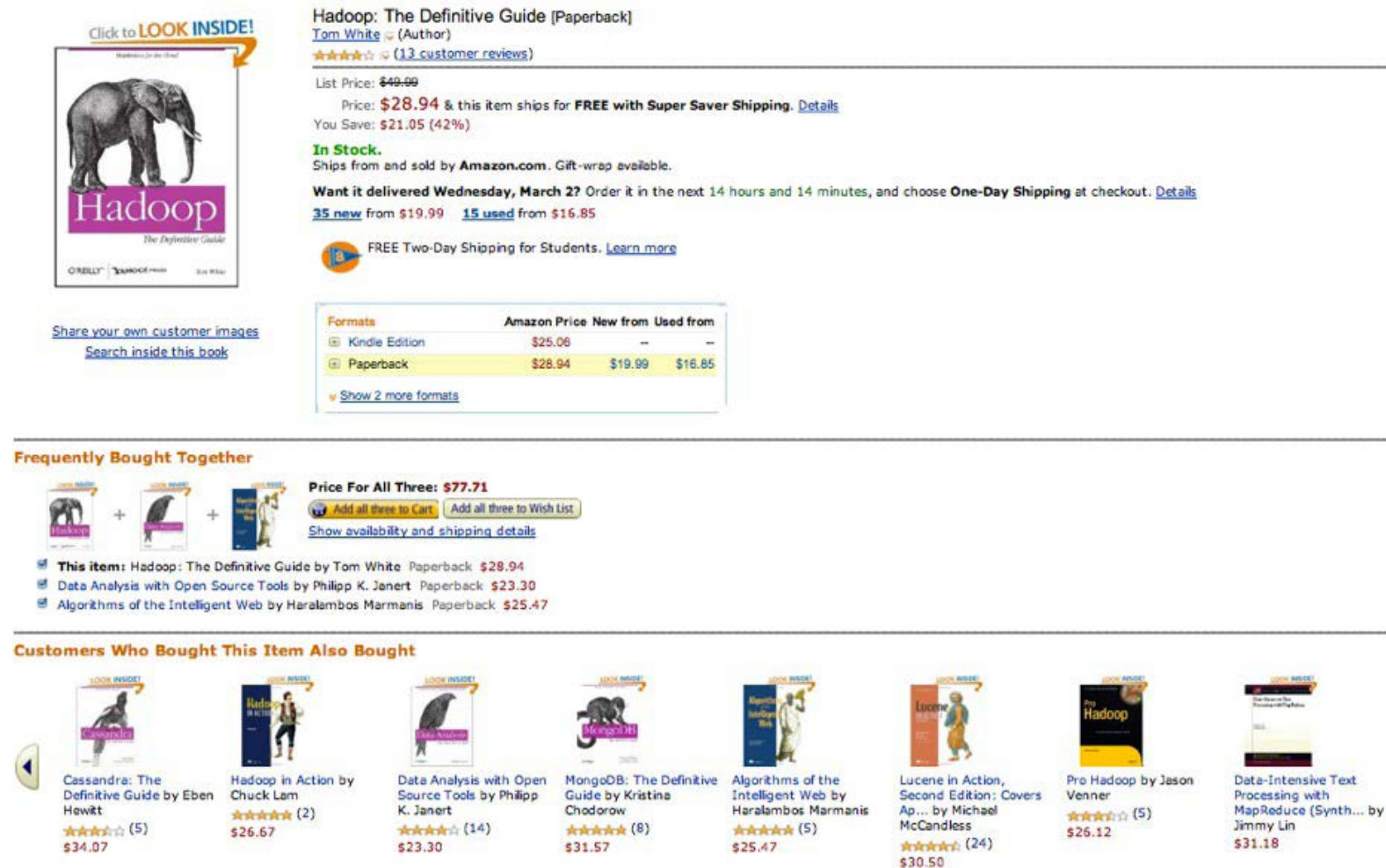
- US government builds their finger print search index using Hadoop

Hadoop Example (2)



- The contents for the *People You May Know* feature is created by a chain of many MapReduce jobs that run daily. The jobs are reportedly a combination of graph traversal, clustering and machine learning

Hadoop Example (3)



The screenshot shows the Amazon product page for 'Hadoop: The Definitive Guide' by Tom White. The page includes a book cover, a 'Click to LOOK INSIDE!' button, and a detailed description. The price is \$28.94, with a list price of \$49.99. The book is in stock and ships for free with Super Saver Shipping. It is available for delivery on Wednesday, March 27. The page also features a 'Frequently Bought Together' section and a 'Customers Who Bought This Item Also Bought' section.


Hadoop: The Definitive Guide (Paperback)
Tom White (Author)
★★★★☆ (13 customer reviews)

List Price: ~~\$49.99~~
Price: **\$28.94** & this item ships for **FREE** with Super Saver Shipping. [Details](#)
You Save: \$21.05 (42%)

In Stock.
Ships from and sold by Amazon.com. Gift-wrap available.

Want it delivered **Wednesday, March 27**? Order it in the next **14 hours and 14 minutes**, and choose **One-Day Shipping** at checkout. [Details](#)

35 new from \$19.99 **15 used** from \$16.85




 **FREE Two-Day Shipping for Students.** [Learn more](#)

[Share your own customer images](#)
[Search inside this book](#)

Formats	Amazon Price	New from	Used from
Kindle Edition	\$25.06	--	--
Paperback	\$28.94	\$19.99	\$16.85

[Show 2 more formats](#)

Frequently Bought Together

 +  +  **Price For All Three: \$77.71**
[Add all three to Cart](#) [Add all three to Wish List](#)
[Show availability and shipping details](#)

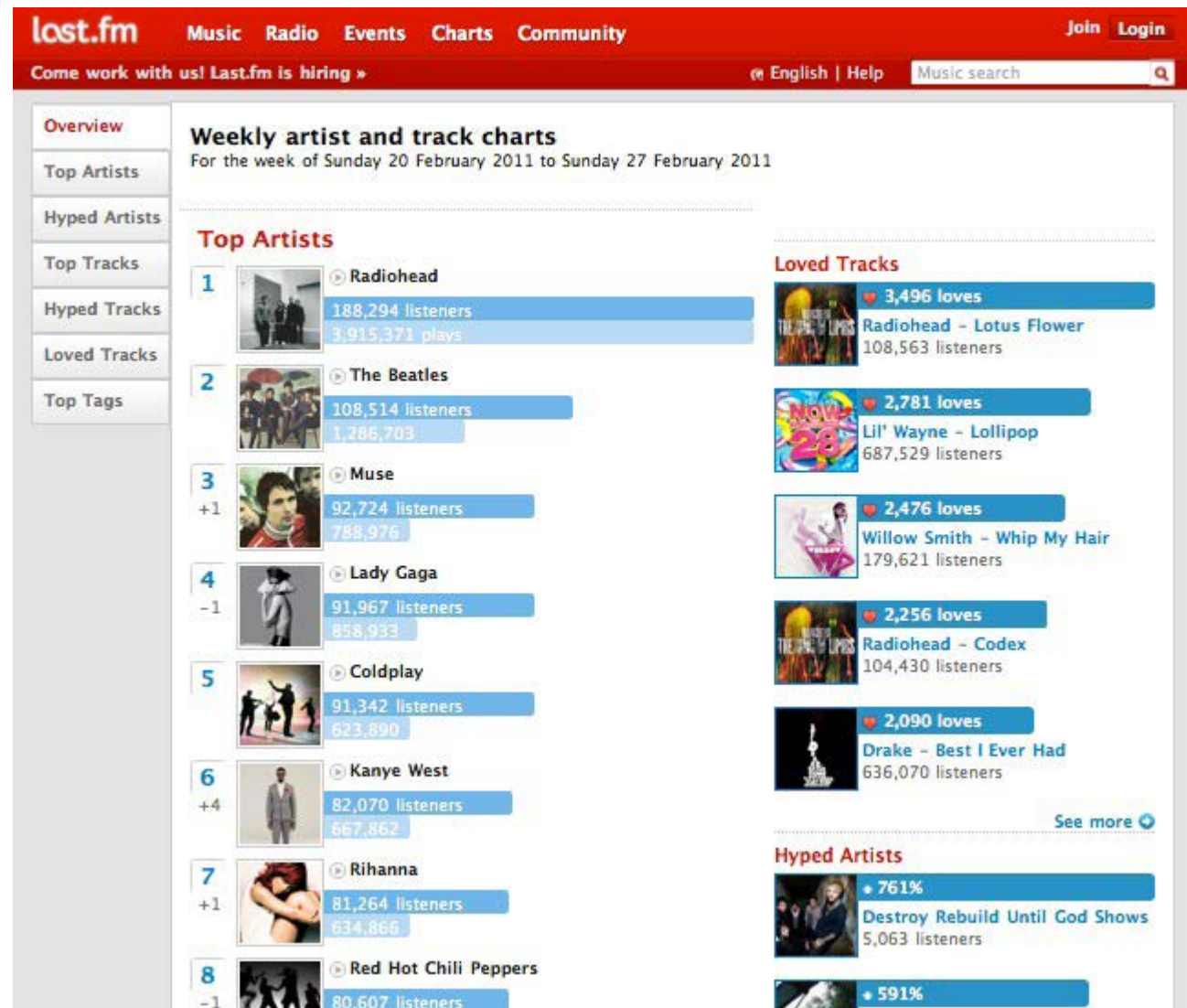
This item: Hadoop: The Definitive Guide by Tom White. Paperback: \$28.94
Data Analysis with Open Source Tools by Philipp K. Janert. Paperback: \$23.30
Algorithms of the Intelligent Web by Haralampos Marmanis. Paperback: \$25.47

Customers Who Bought This Item Also Bought

Book Title	Author	Price	Reviews
Cassandra: The Definitive Guide	Eben Hewitt	\$34.07	★★★★☆ (5)
Hadoop in Action	Chuck Lam	\$26.67	★★★★★ (2)
Data Analysis with Open Source Tools	Philipp K. Janert	\$23.30	★★★★☆ (14)
MongoDB: The Definitive Guide	Kristina Chodorow	\$31.57	★★★★★ (8)
Algorithms of the Intelligent Web	Haralampos Marmanis	\$25.47	★★★★★ (5)
Lucene in Action, Second Edition: Covers Apache Lucene 4.x	Michael McCandless	\$30.50	★★★★★ (24)
Pro Hadoop	Jason Venner	\$26.12	★★★★☆ (5)
Data-Intensive Text Processing with MapReduce	Jimmy Lin	\$31.18	

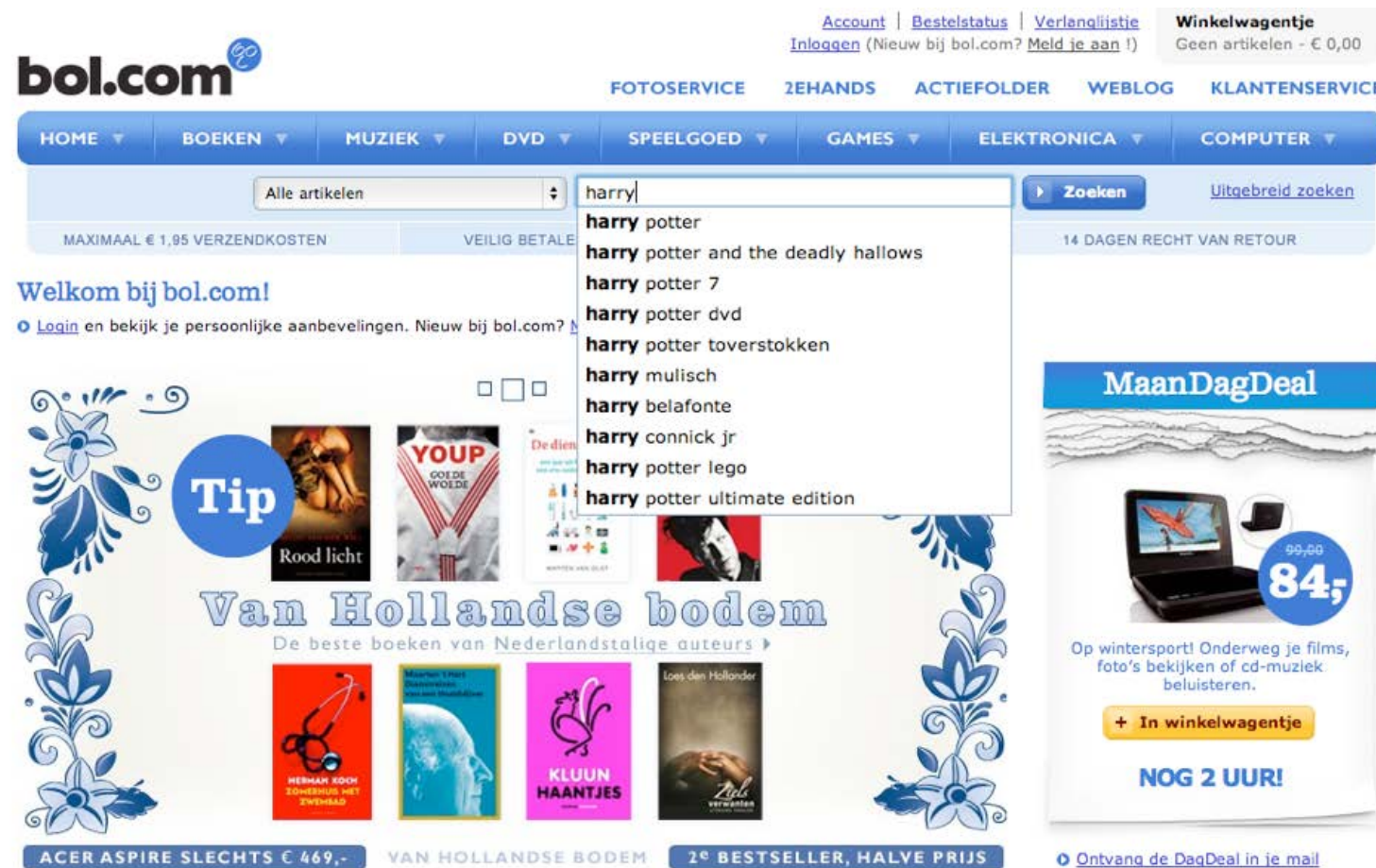
- Amazon's *Frequently Bought Together* and *Customers Who Bought This Item Also Bought* features are brought to you by MapReduce jobs.

Hadoop Example (4)



- Top Charts generated daily based on millions of users' listening behavior.

Hadoop Example (5)

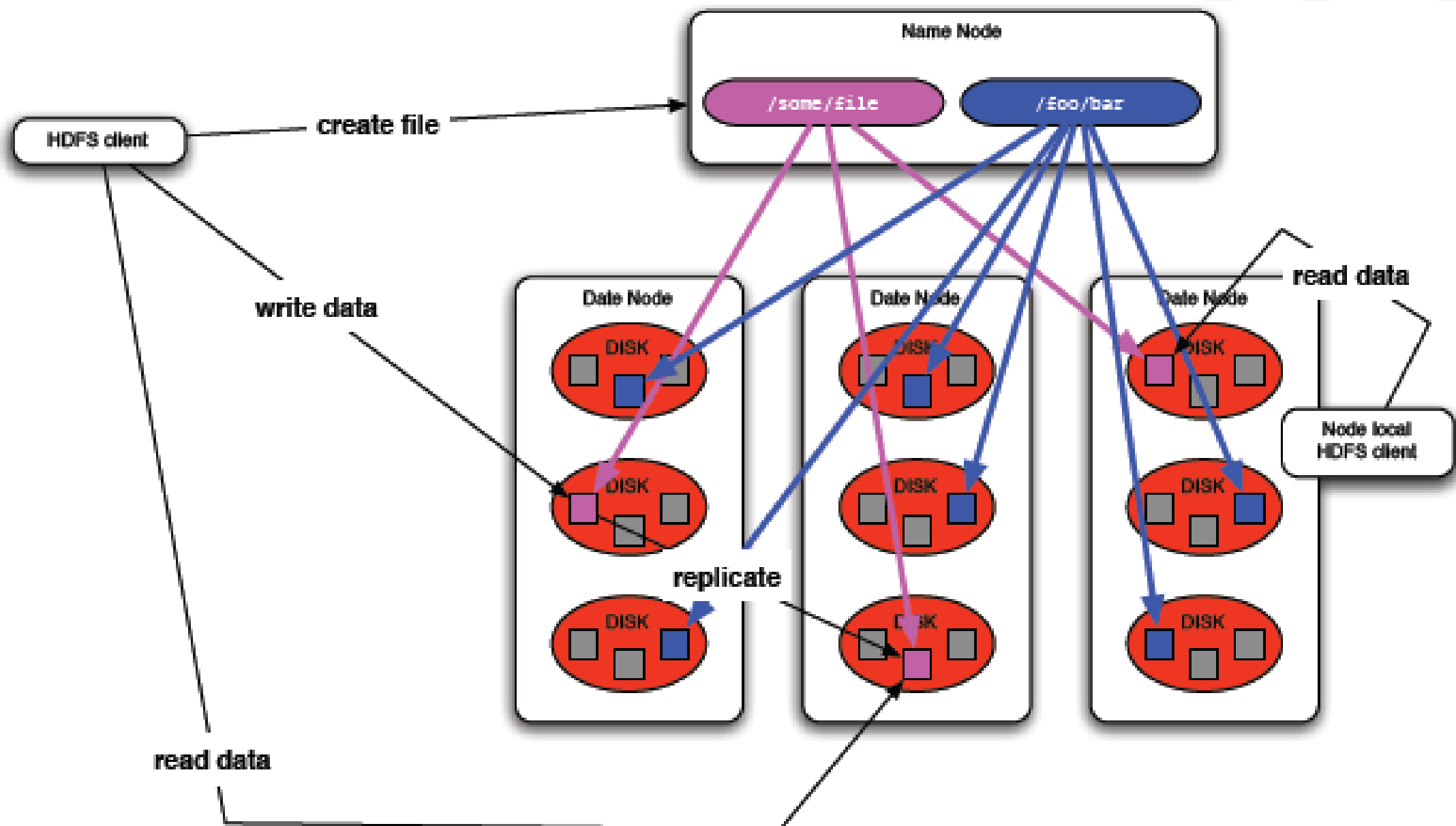


- Top searches used for auto-completion are re-generated daily by a MapReduce job using all searches for the past couple of days. Popularity for search terms can be based on counts, but also trending and correlation with other datasets (e.g. trending on social media, news, charts in case of music and movies, best seller lists, etc.)

Hadoop File System (1)

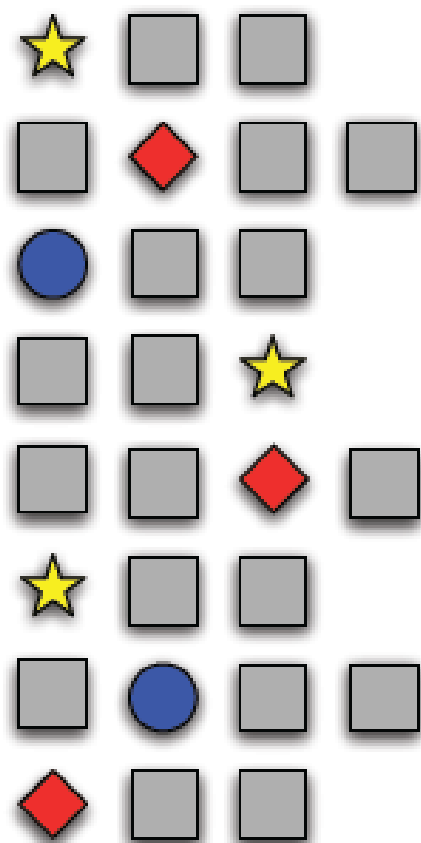
- Distributed file system
- Consists of a single master node and multiple (many) data nodes
- Files are split up in blocks (typically 64MB)
- Blocks are spread across data nodes in the cluster
- Each block is replicated multiple times to different data nodes in the cluster (typically 3 times)
- Master node keeps track of which blocks belong to a file

Hadoop File System (2)

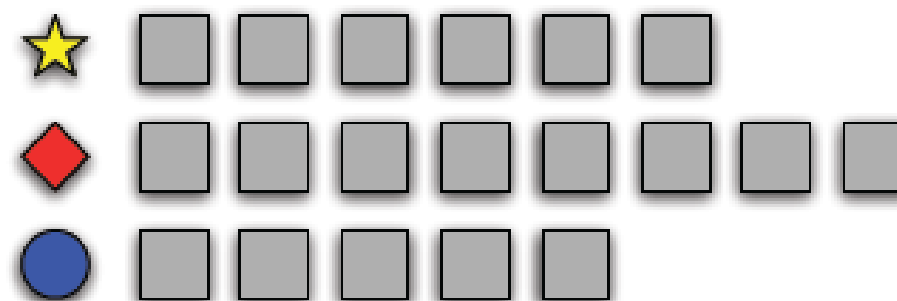


MapReduce

Input data

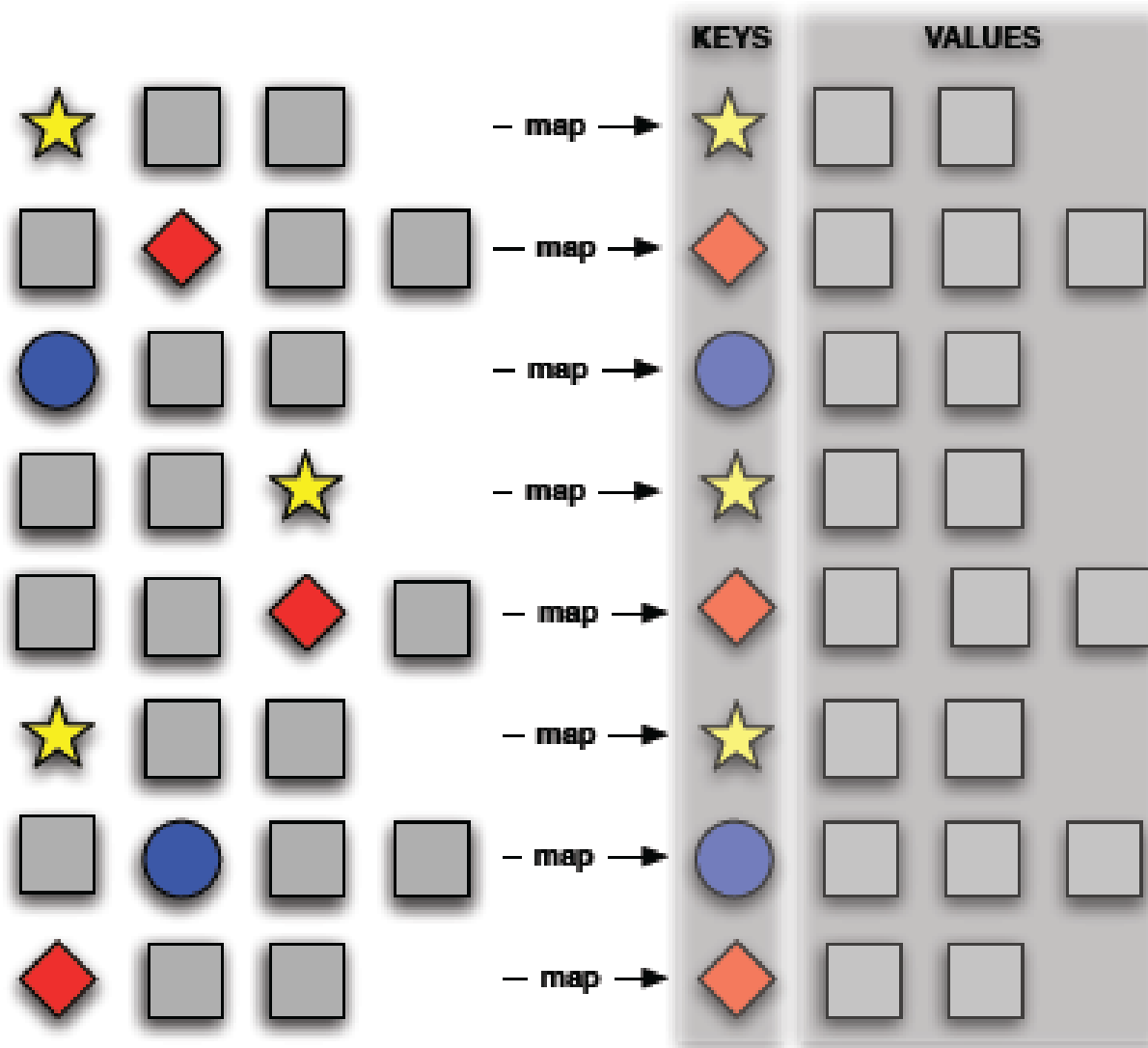


Required output



Map Step

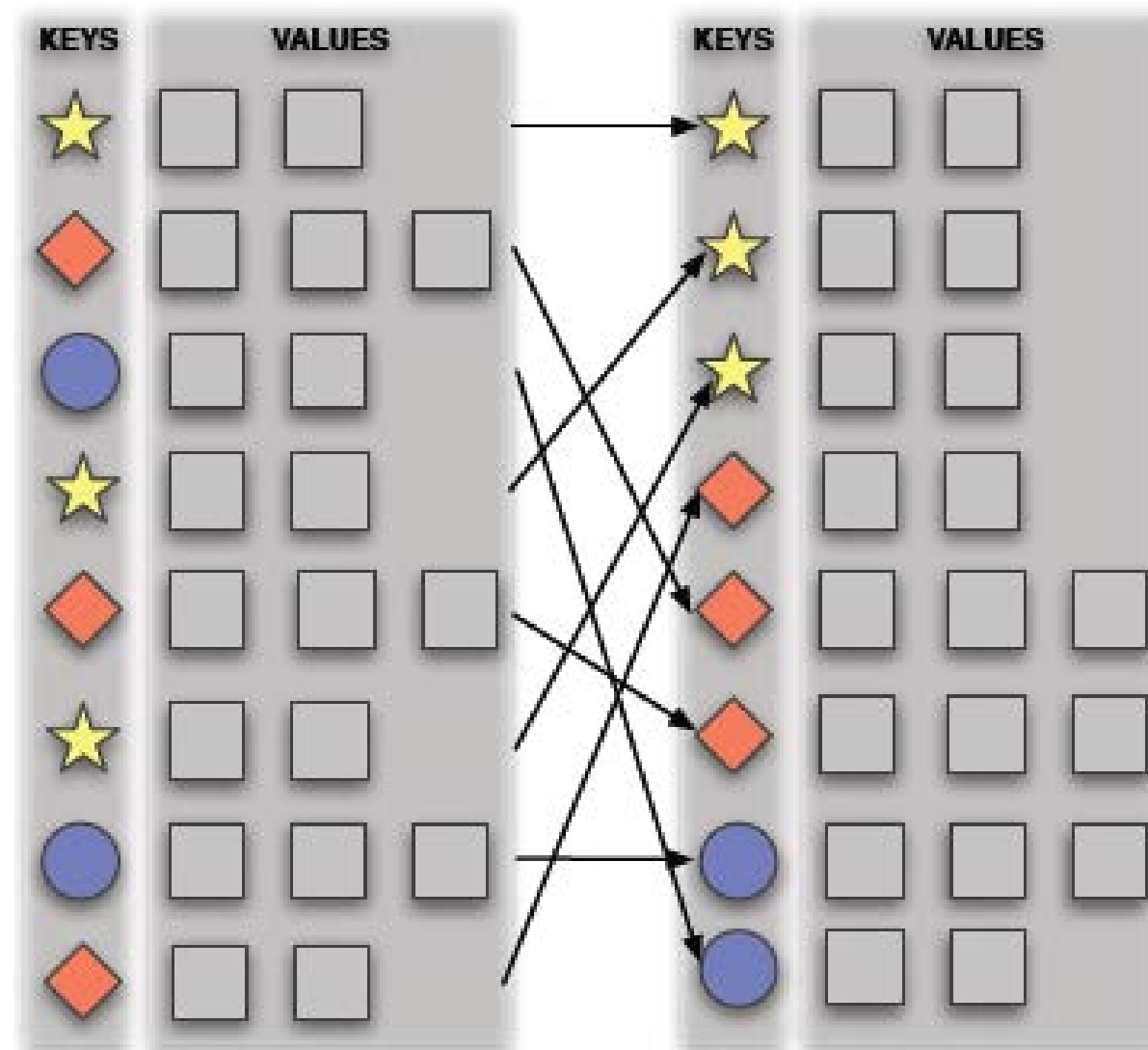
- Map: extract something useful from each record



```
void map(recordNumber, record) {  
    key = record.findColorfulShape();  
    value = record.findGrayShapes();  
    emit(key, value);  
}
```

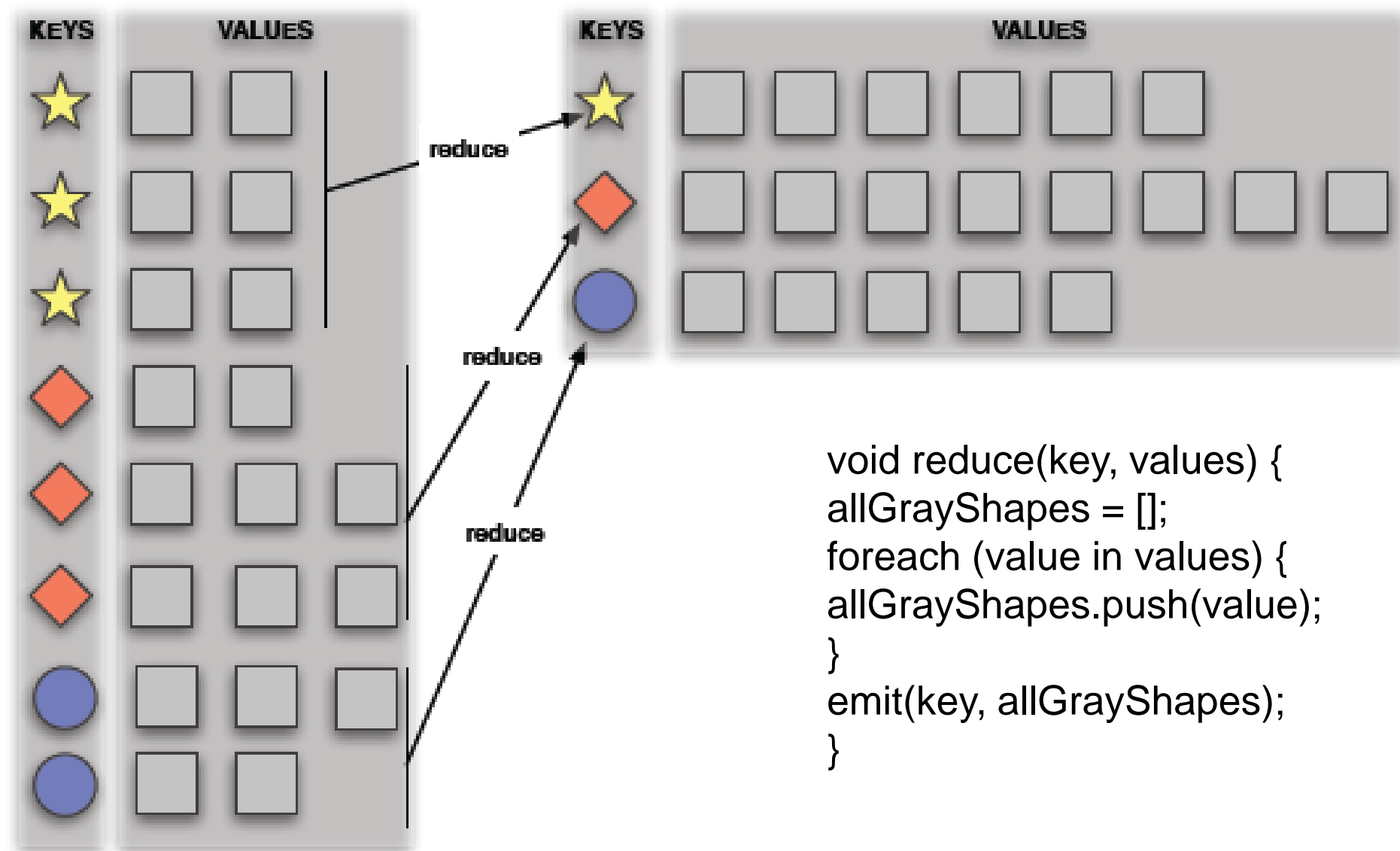

Sorting Keys

- Framework sorts all KeyValue pairs by Key

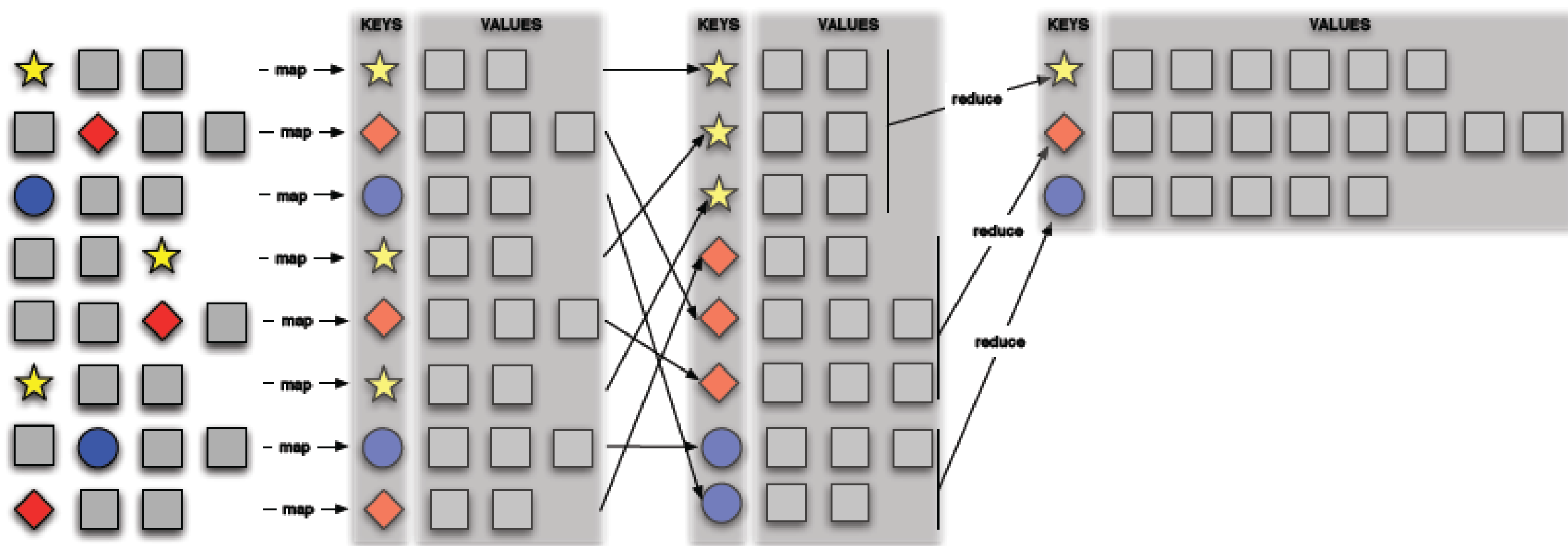


Reduce Step

- Reduce: process values for each key



MapReduce



Hadoop / MapReduce for Data Mining

- Many machine learning / data mining algorithms can be parallelized, i.e., rewritten into Map and Reduce steps
- Apache Mahout: scalable machine learning library

