Data

BT 3041: Analysis and interpretation of Biological Data

Outline

- Types of data
- Data visualisation

Types of Data

- **Data Set** is a collection of **Data Objects**
- Data Object: also called
  - Record, point, vector, pattern, event, case, sample, observation, entity
- Data object is represented by a set of **Attributes**
- **Attribute:** is a property or a characteristic of an object

Types of attributes

- Qualitative
  - Nominal: just a name.
    - eg. pincodes, IDs, eye color
  - Ordinal: information to order objects.
    - Eg. {good, better, best}
- Quantitative
  - Numerical value exists
    - Eg. temperature, pH

Quantitative attributes

- Discrete
  - Binary
- Continuous

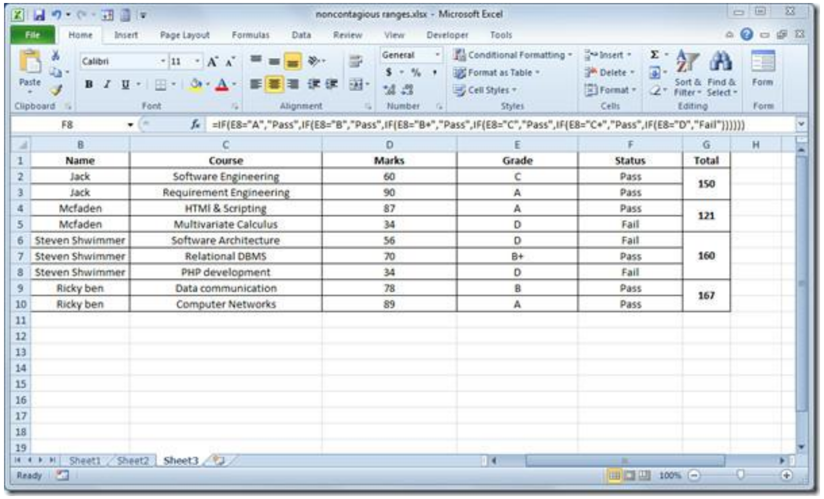
Characteristics of Data sets

- Dimensionality
  - Curse of dimensionality
  - Dimensionality reduction
- Sparsity
  - Only a small fraction of attributes are non-zero
- Resolution
  - Converting continuous quantities to discrete ones

Types of data sets

- Record Data
- Graph/network Data
- Ordered data
- Spatial, image and multimedia data

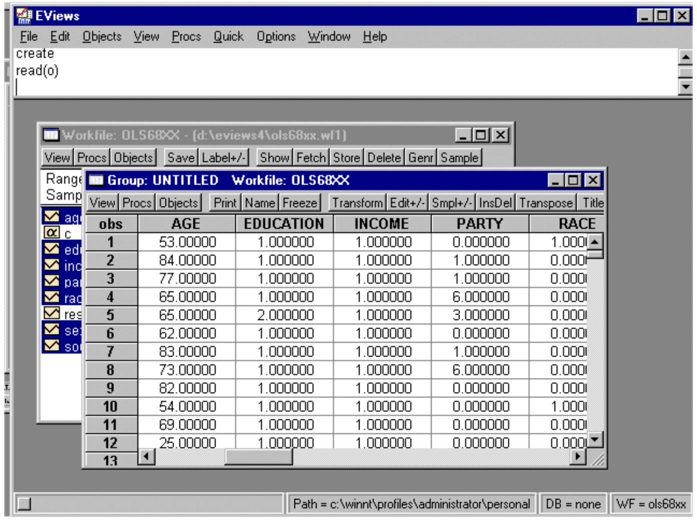
Record data:
Same number of attributes



F8 =IF(E8="A","Pass",IF(E8="B","Pass",IF(E8="B+","Pass",IF(E8="C","Pass",IF(E8="C+","Pass",IF(E8="D","Fail"))))))

	B	C	D	E	F	G
1	Name	Course	Marks	Grade	Status	Total
2	Jack	Software Engineering	60	C	Pass	150
3	Jack	Requirement Engineering	90	A	Pass	
4	Mcfaden	HTMl & Scripting	87	A	Pass	121
5	Mcfaden	Multivariate Calculus	34	D	Fail	
6	Steven Shwimmer	Software Architecture	56	D	Fail	160
7	Steven Shwimmer	Relational DBMS	70	B+	Pass	
8	Steven Shwimmer	PHP development	34	D	Fail	
9	Ricky ben	Data communication	78	B	Pass	167
10	Ricky ben	Computer Networks	89	A	Pass	

Data matrix



```
create
read(o)
```

obs	AGE	EDUCATION	INCOME	PARTY	RACE
1	53.00000	1.000000	1.000000	0.000000	1.000
2	84.00000	1.000000	1.000000	1.000000	0.000
3	77.00000	1.000000	1.000000	1.000000	0.000
4	65.00000	1.000000	1.000000	6.000000	0.000
5	65.00000	2.000000	1.000000	3.000000	0.000
6	62.00000	1.000000	1.000000	0.000000	0.000
7	83.00000	1.000000	1.000000	1.000000	0.000
8	73.00000	1.000000	1.000000	6.000000	0.000
9	82.00000	1.000000	1.000000	0.000000	0.000
10	54.00000	1.000000	1.000000	0.000000	1.000
11	69.00000	1.000000	1.000000	0.000000	0.000
12	25.00000	1.000000	1.000000	0.000000	0.000
13					

Document data

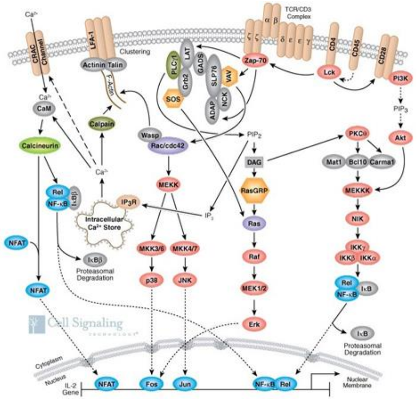
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Market basket data

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

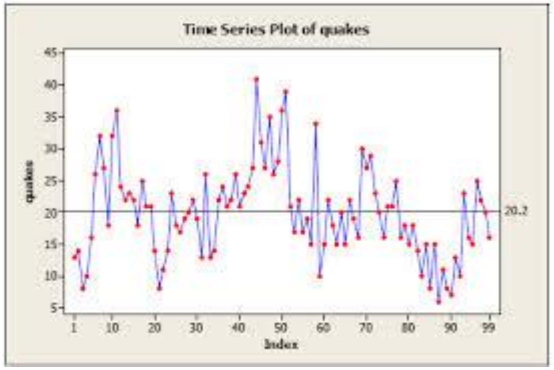
Graph/Network data

- Examples:
  - Internet
  - Signaling networks
  - Molecular structures

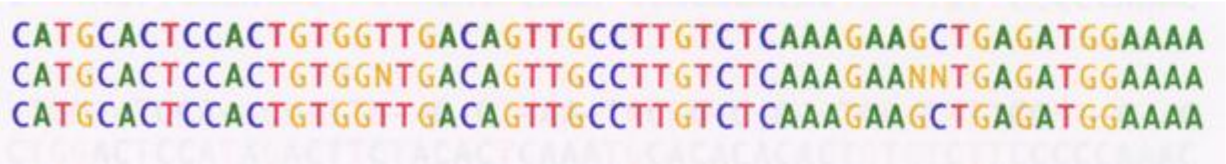


Ordered Data

- Video data
- Time series data



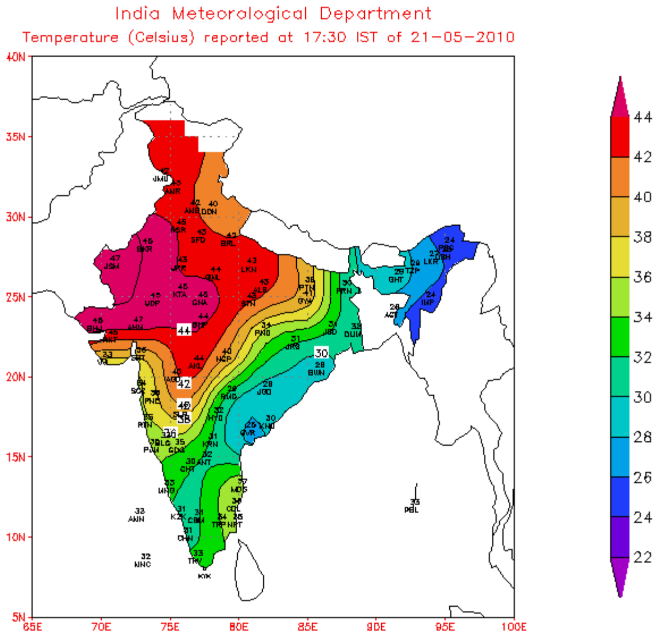
- Molecular sequence data



```
CATGCACTCCACTGTGGTTGACAGTTGCCTTGTCTCAAAGAAGCTGAGATGGAAAA
CATGCACTCCACTGTGGNTGACAGTTGCCTTGTCTCAAAGAANNTGAGATGGAAAA
CATGCACTCCACTGTGGTTGACAGTTGCCTTGTCTCAAAGAAGCTGAGATGGAAAA
```

Spatial/image data

- Examples:
  - Maps
  - Images



India Meteorological Department
Temperature (Celsius) reported at 17:30 IST of 21-05-2010
40N
35N
30N
25N
20N
15N
10N
5N
65E
70E
75E
80E
85E
90E
95E
100E
44
42
40
38
36
34
32
30
28
26
24
22

http://imgarcade.com/1/india-temperature-map/

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.

(Hans, Kamber, Pei 2013)

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

  Note: *n* is sample size and *N* is population size.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

  – Weighted arithmetic mean:

  – Trimmed mean: chopping extreme values

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Median:

  – Middle value if odd number of values, or average of the middle two values otherwise

	age	*frequency*
	1–5	200
	6–15	450
	16–20	300
Median interval →	21–50	1500
	51–80	700
	81–110	44

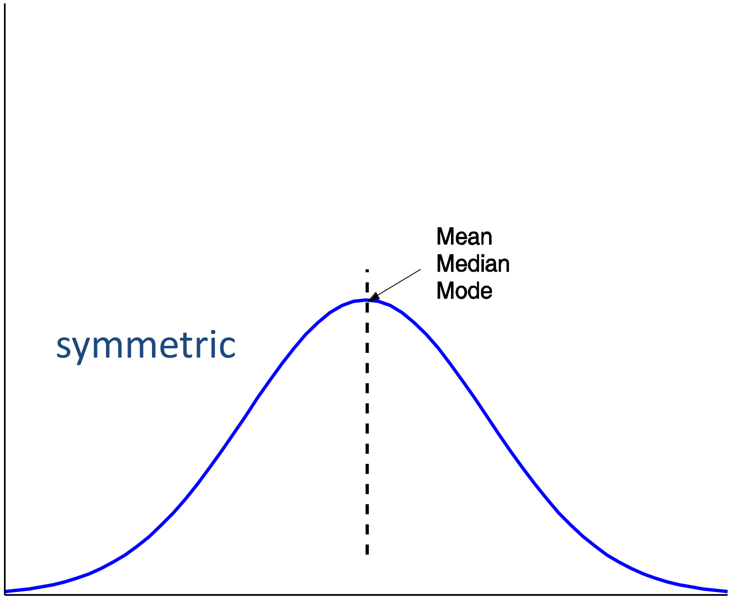
- Mode

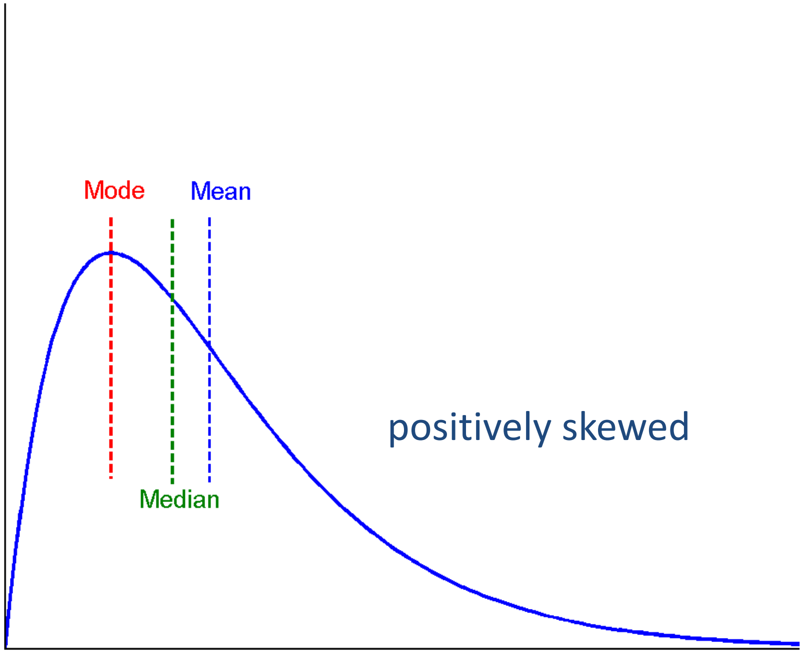
  – Value that occurs most frequently in the data

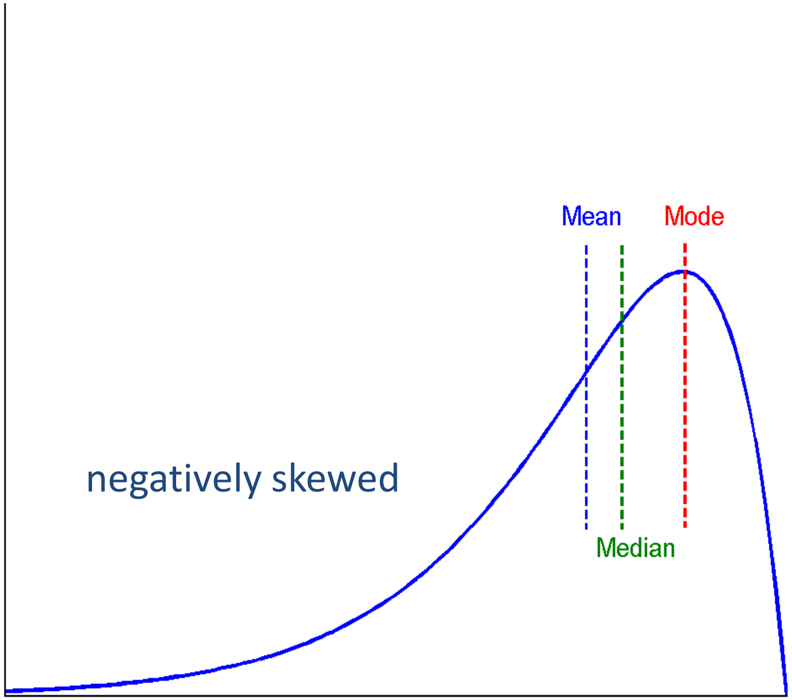
  – Unimodal, bimodal, trimodal

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







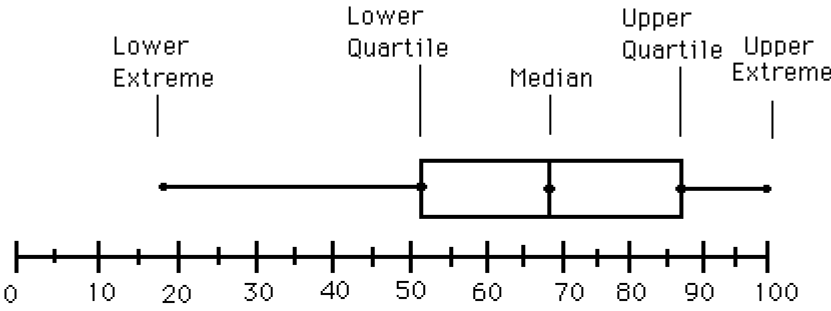
Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**:
    - Q_1 (25th percentile): The first **quartile** (Q_1) is defined as the middle number between the smallest number and the median of the data set.
    - The second **quartile** (Q_2) is the median of the data.
    - Q_3 (75th percentile): The third **quartile** (Q_3) is the middle value between the median and the highest value of the data set.
  - **Inter-quartile range**: $IQR = Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max

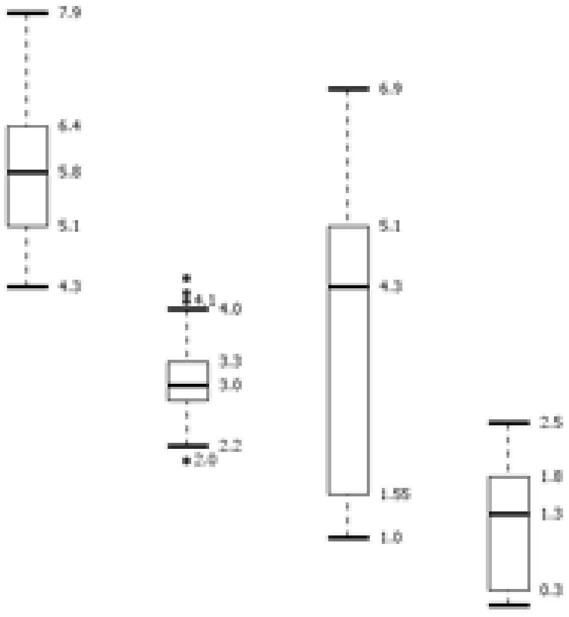
Boxplot

- **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
- **Outlier**: usually, a value higher/lower than 1.5 x IQR

Boxplot Analysis

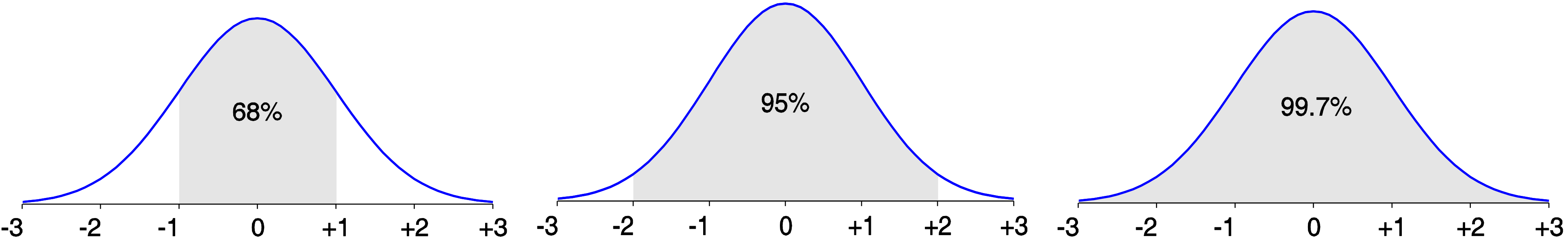


- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually

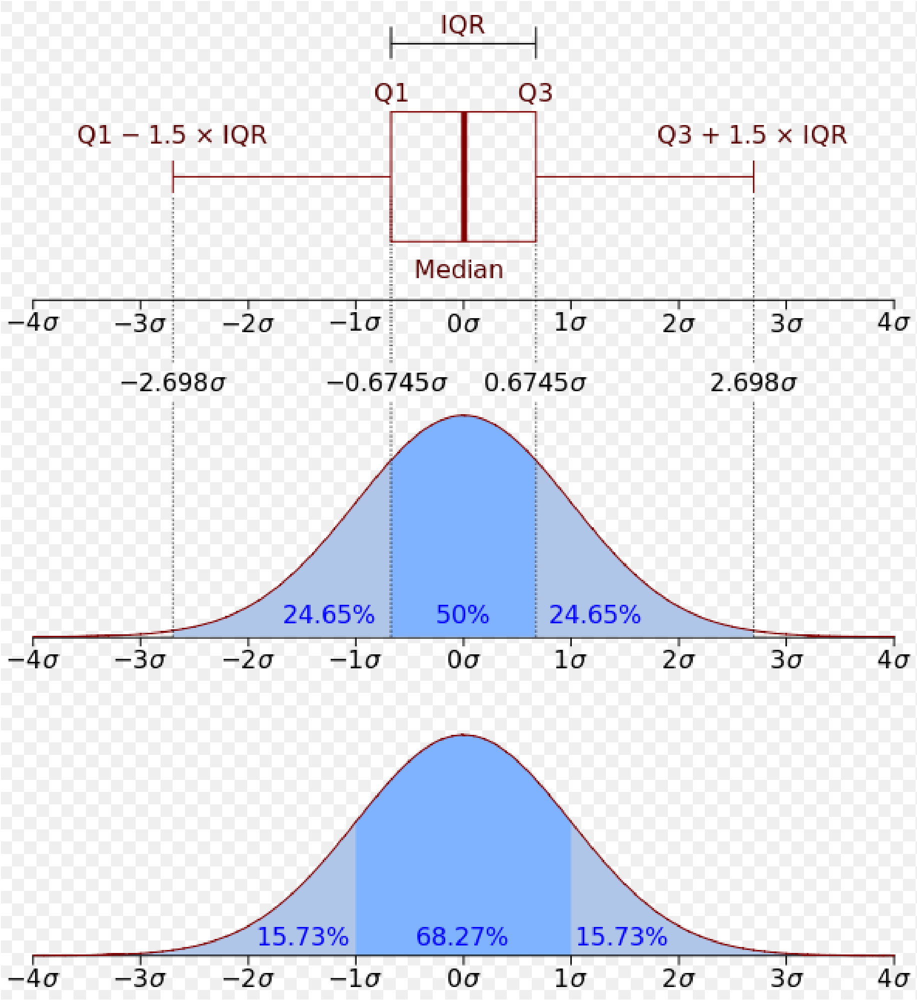


Properties of Normal Distribution Curve

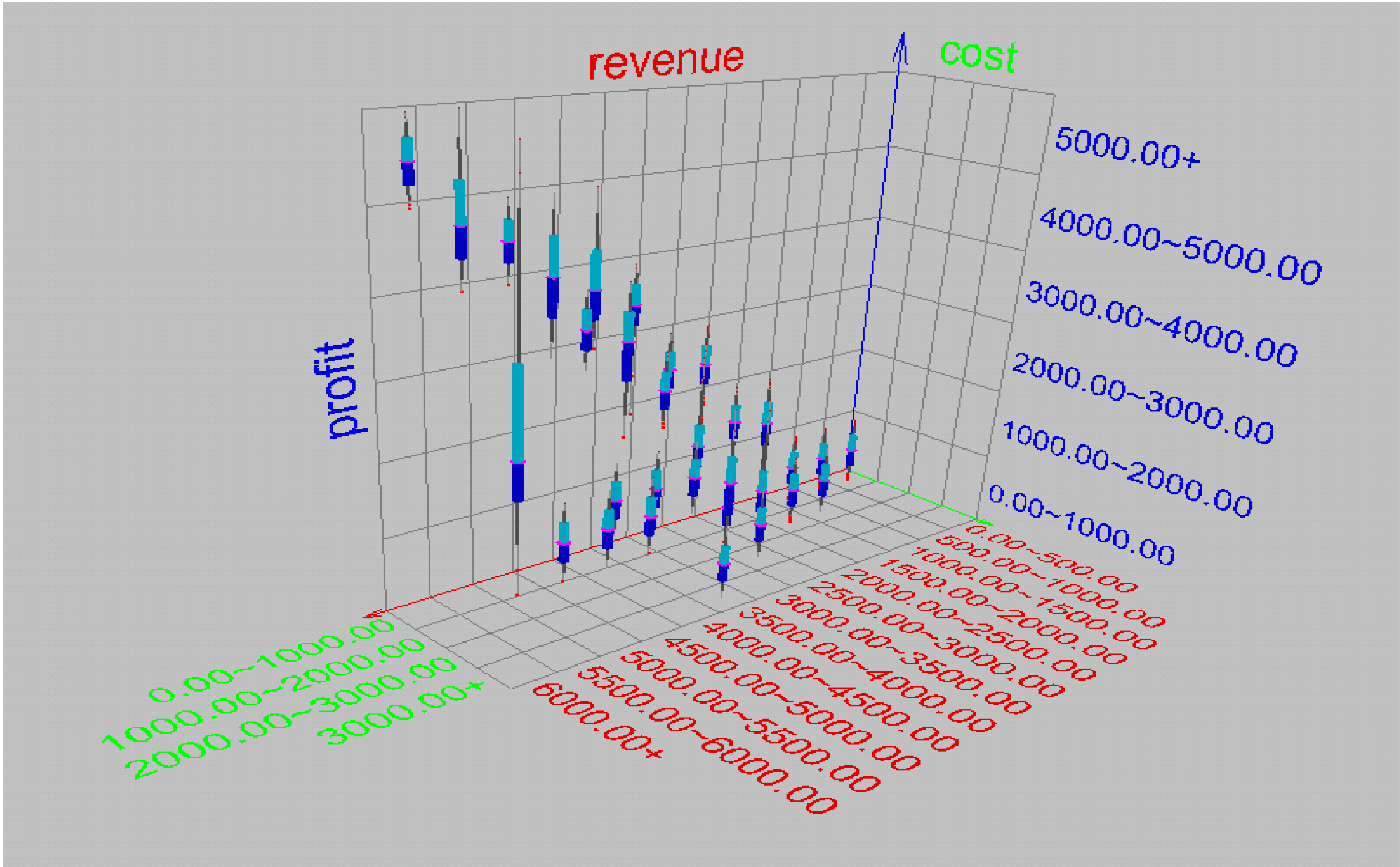
- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it



Boxplot for normal distribution



Visualization of Data Dispersion: 3-D Boxplots



Variance

- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^{\ 2} - \mu^2$$

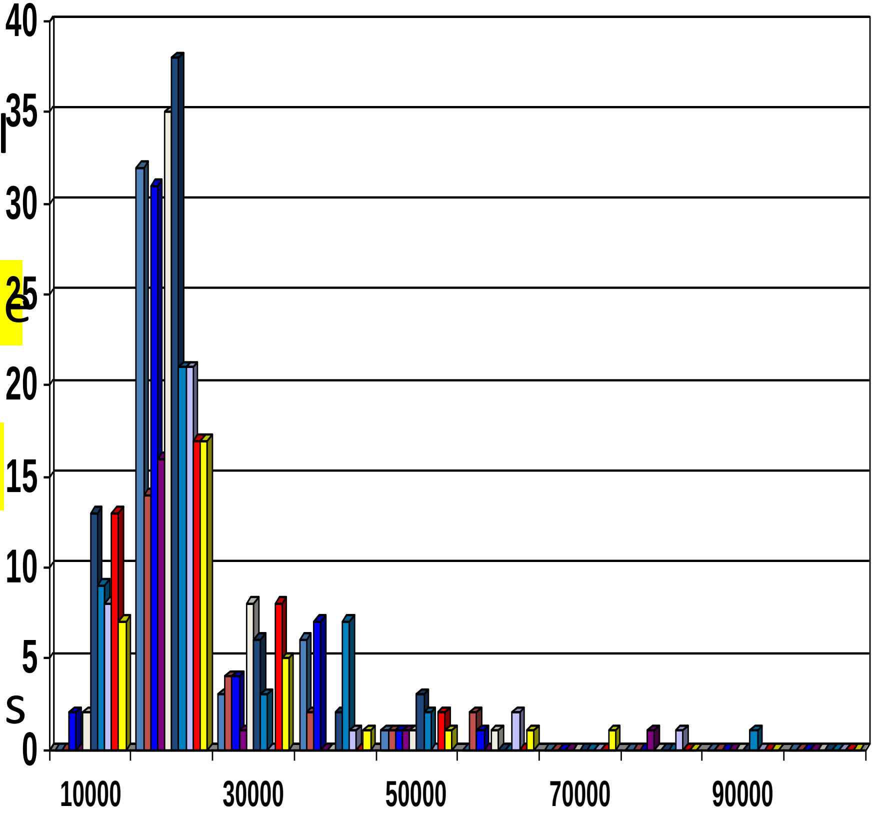
  - **Standard deviation** *σ* is the square root of variance *or σ*^{2}

Graphic Displays of Basic Statistical Descriptions

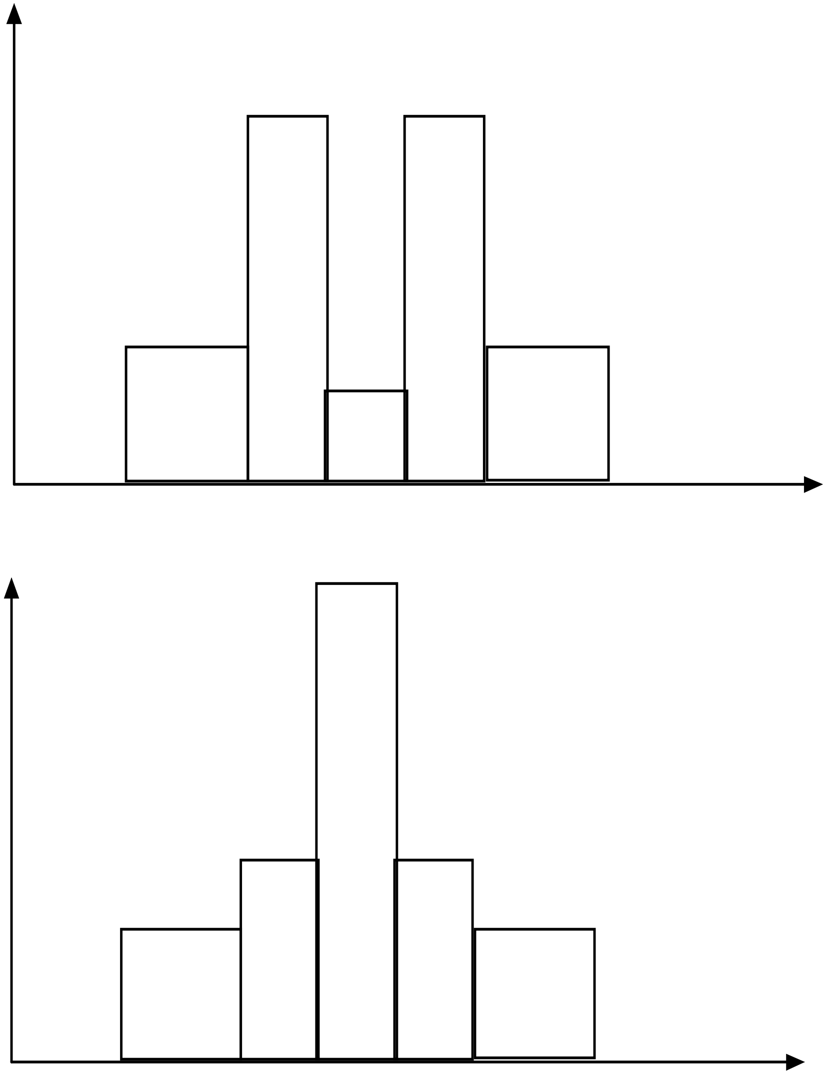
- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



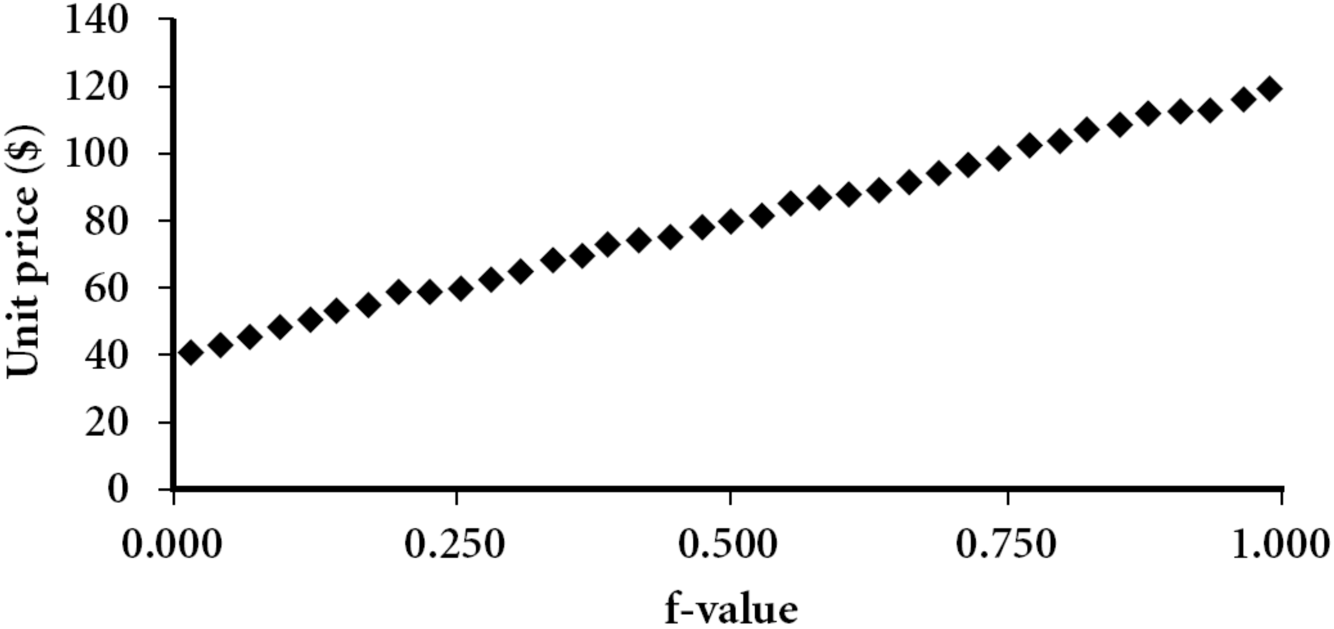
Histograms Often Tell More than Boxplots



- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

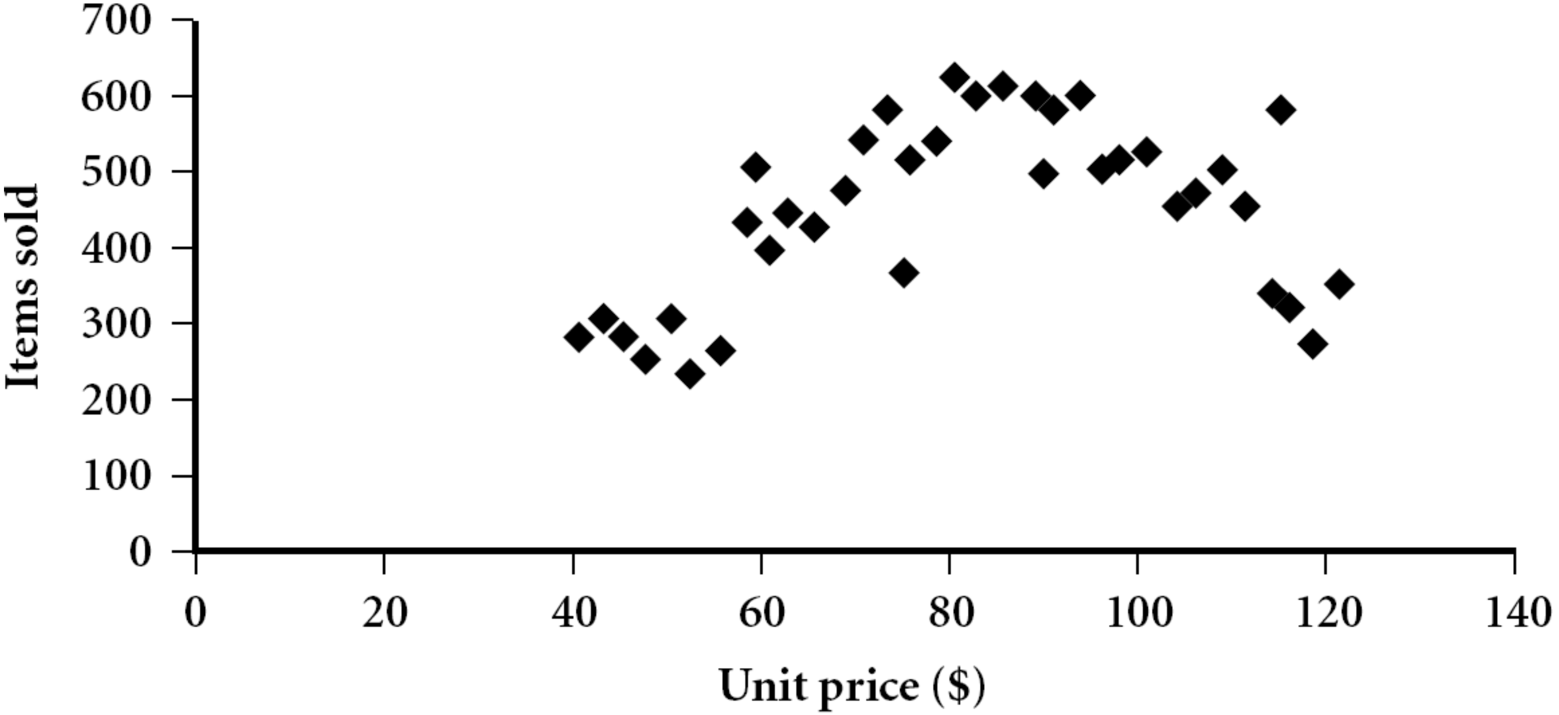
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Data Visualization

Data Visualization

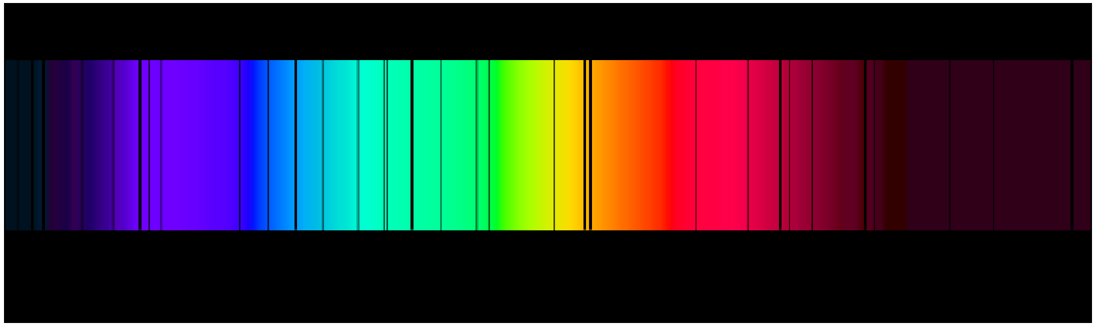
- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters for further quantitative analysis

VISUALIZING IN LOW DIMENSIONS

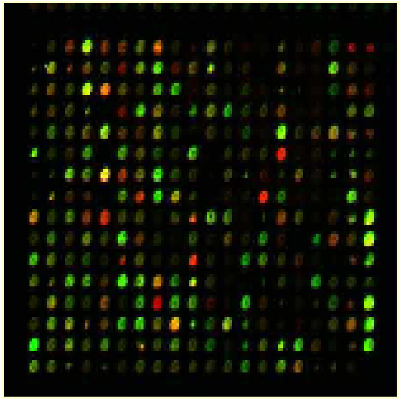
Pixel-Oriented Visualization Techniques

- 1D

Solar spectrum

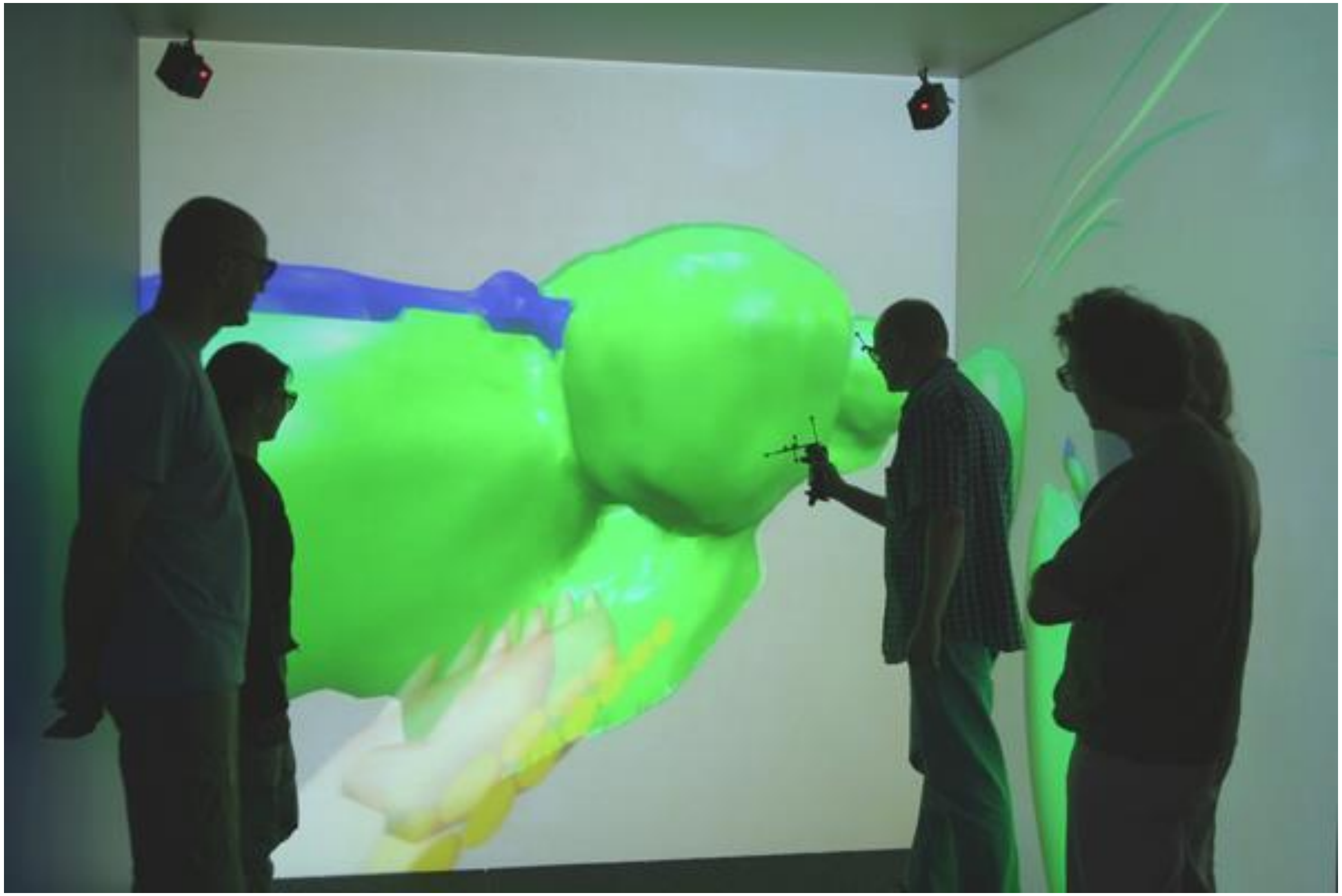


- 2D



Microarray data

3D visualization



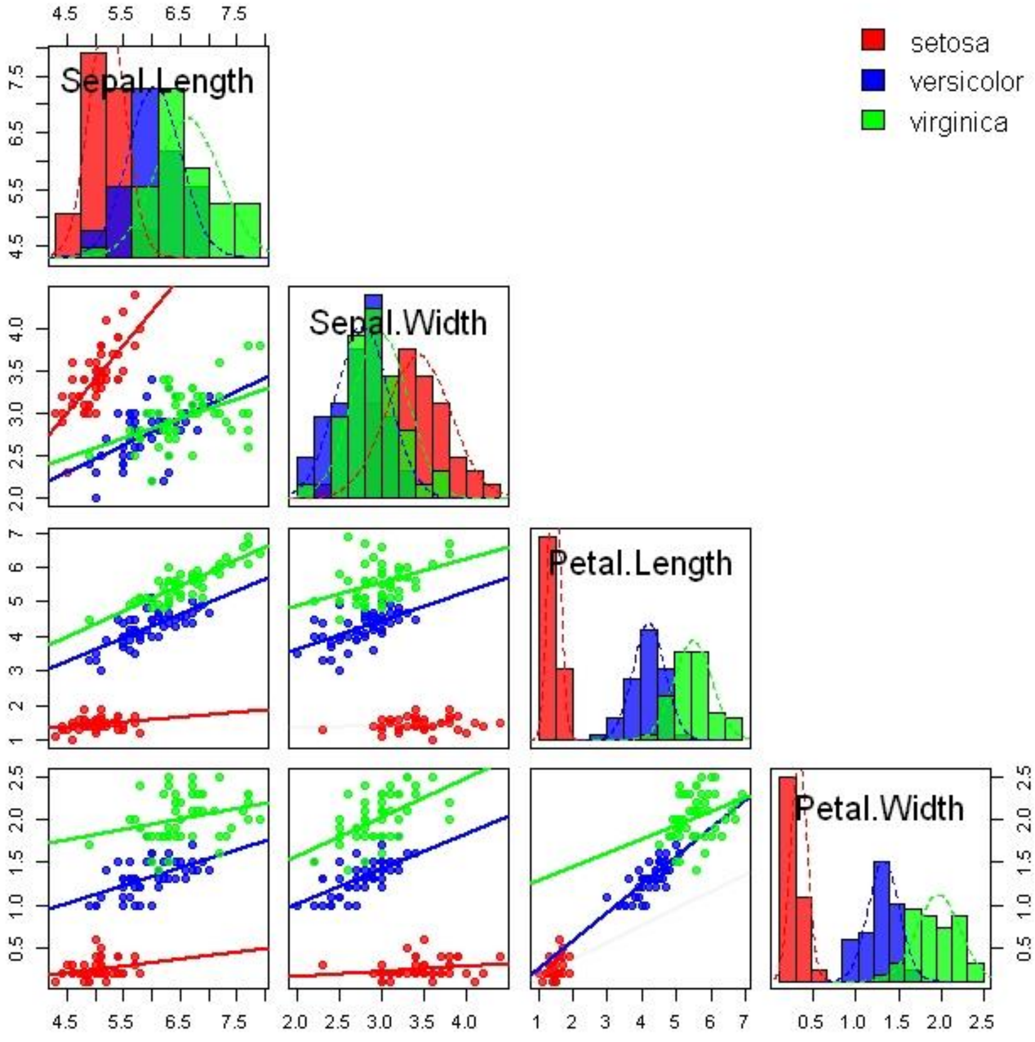
Interactive visualization of three-dimensional biological models in a CAVE.

http://www.jvrb.org/past-issues/6.2009/2257

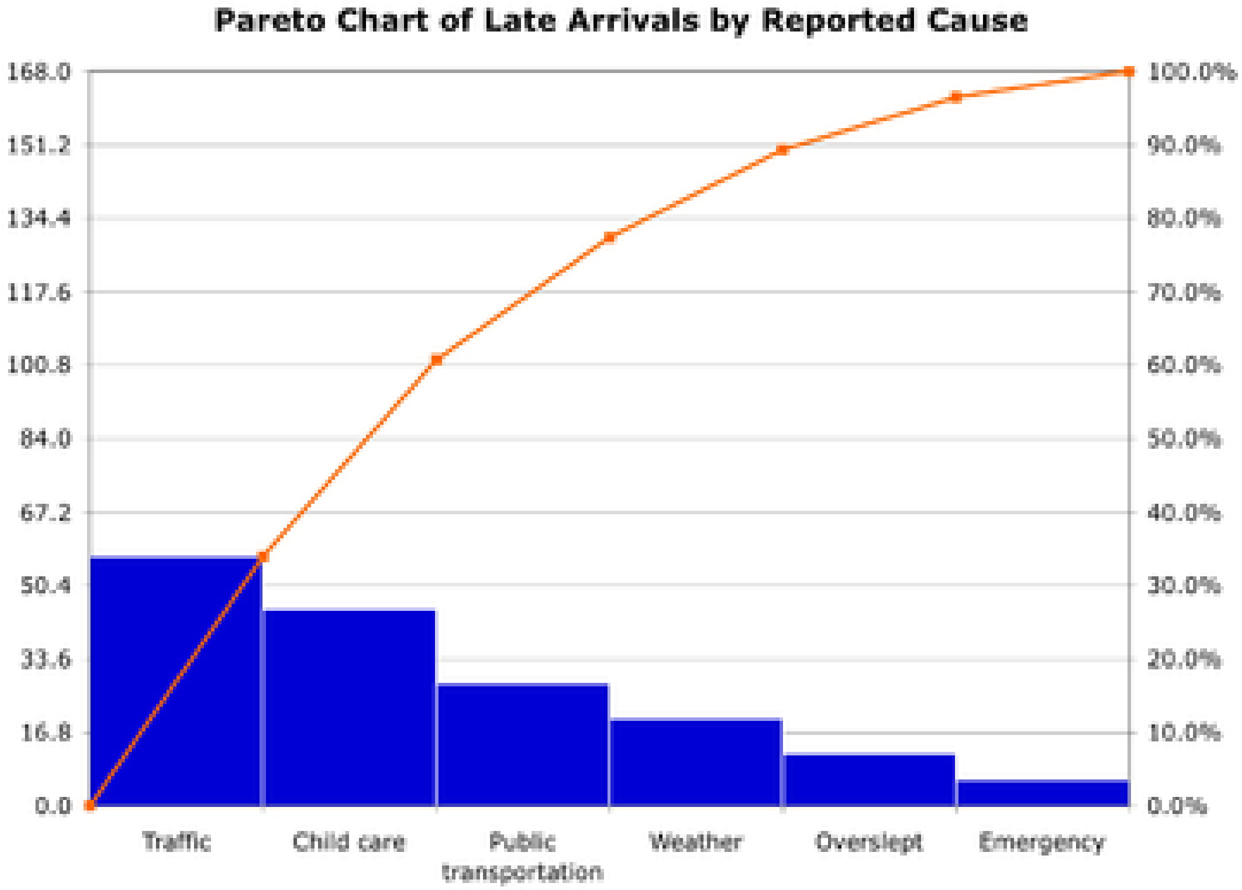
Histogram

- Ex: Iris data
- 5 kinds of Iris flowers
- 4 attributes: petal length/width, sepal length/width
- 50 samples per each flower type

Iris data: histograms of individual attributes

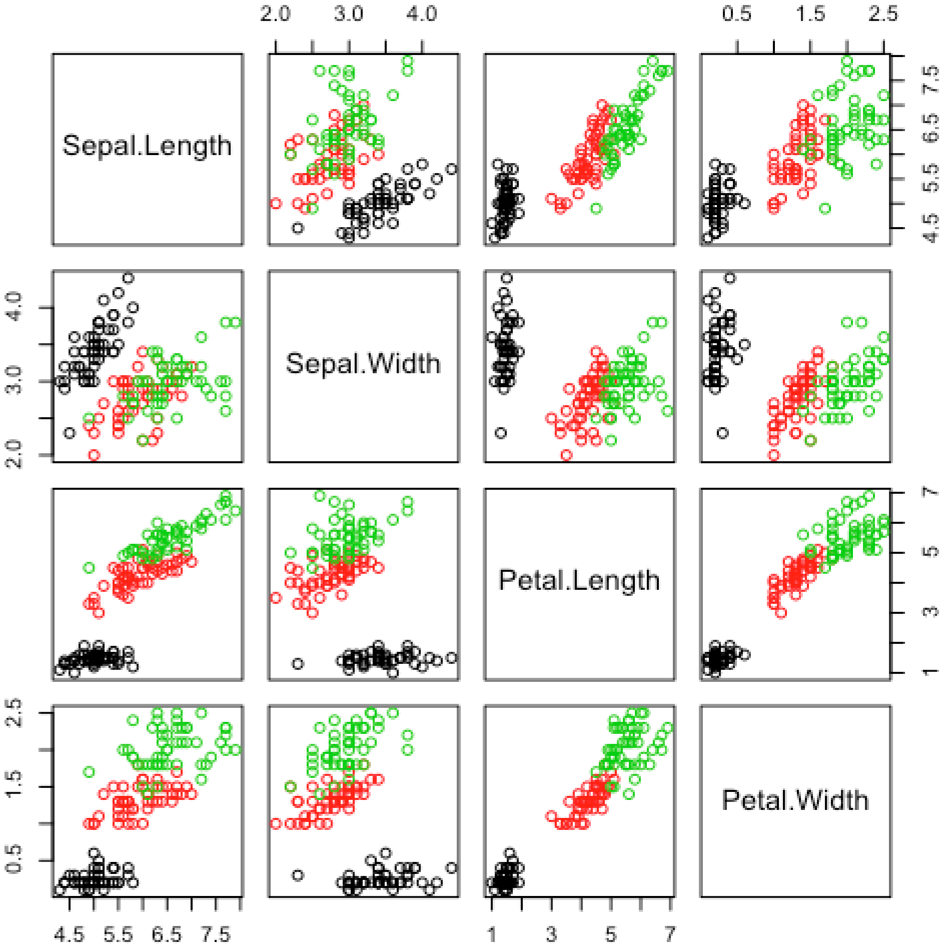


Pareto histogram

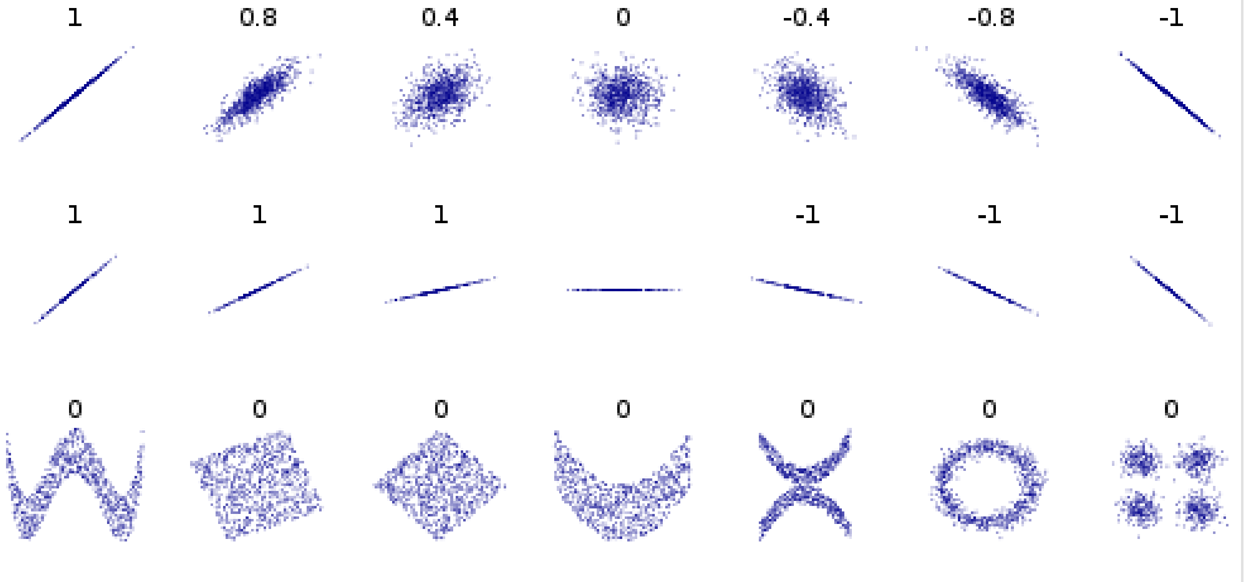


Frequencies are sorted

Scatter plots

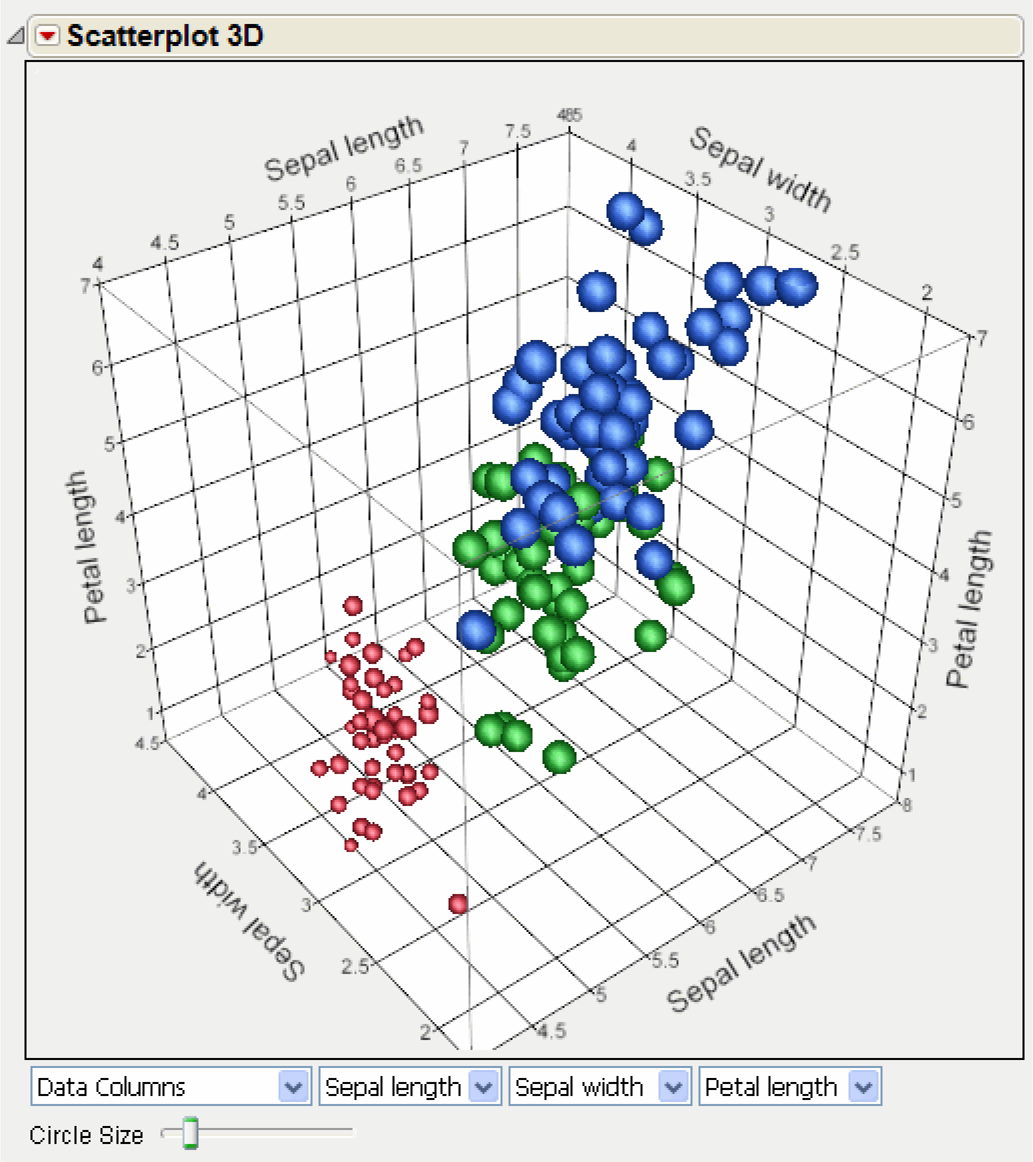


Correlation from scatter plots



$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

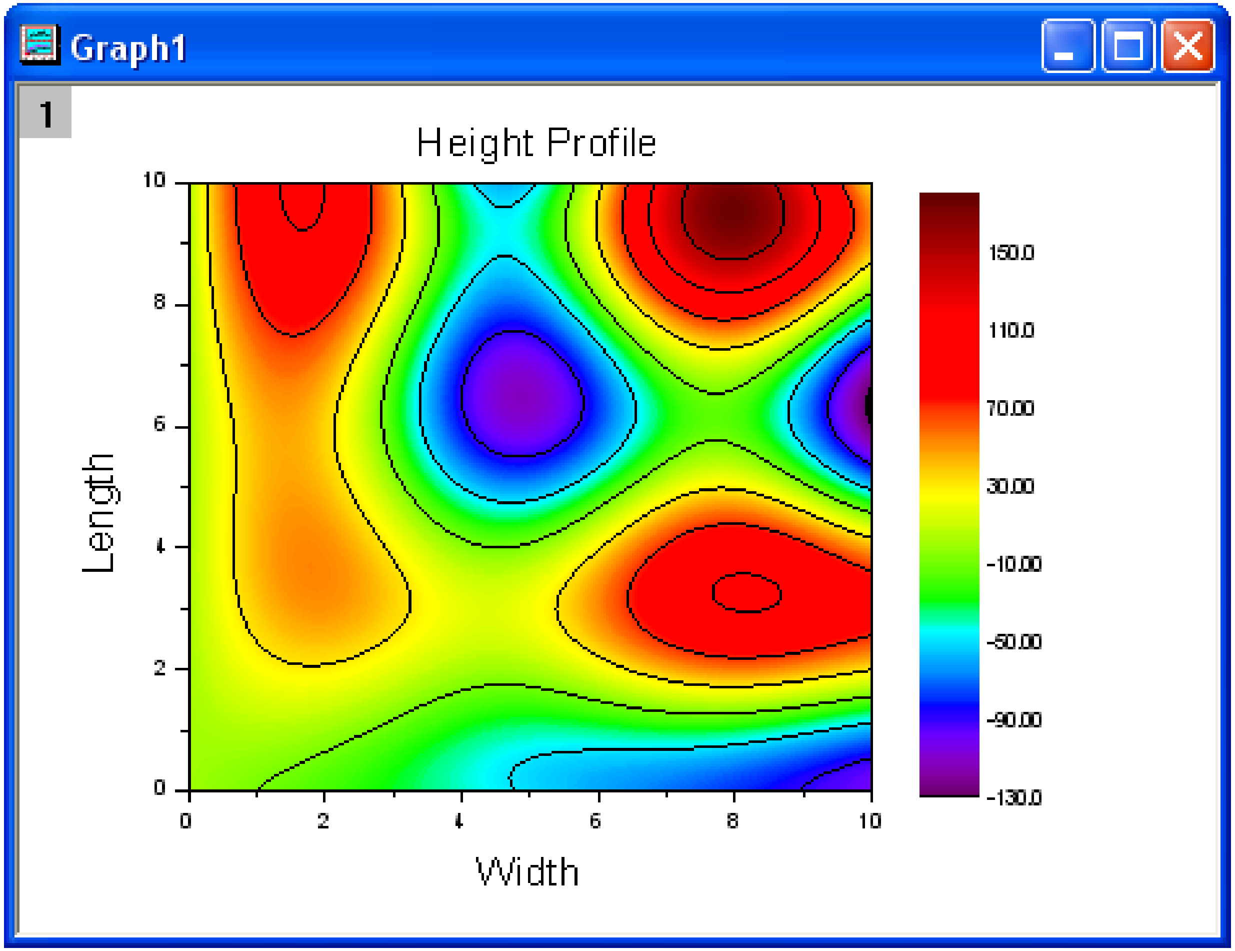
Scatter plot in 3D



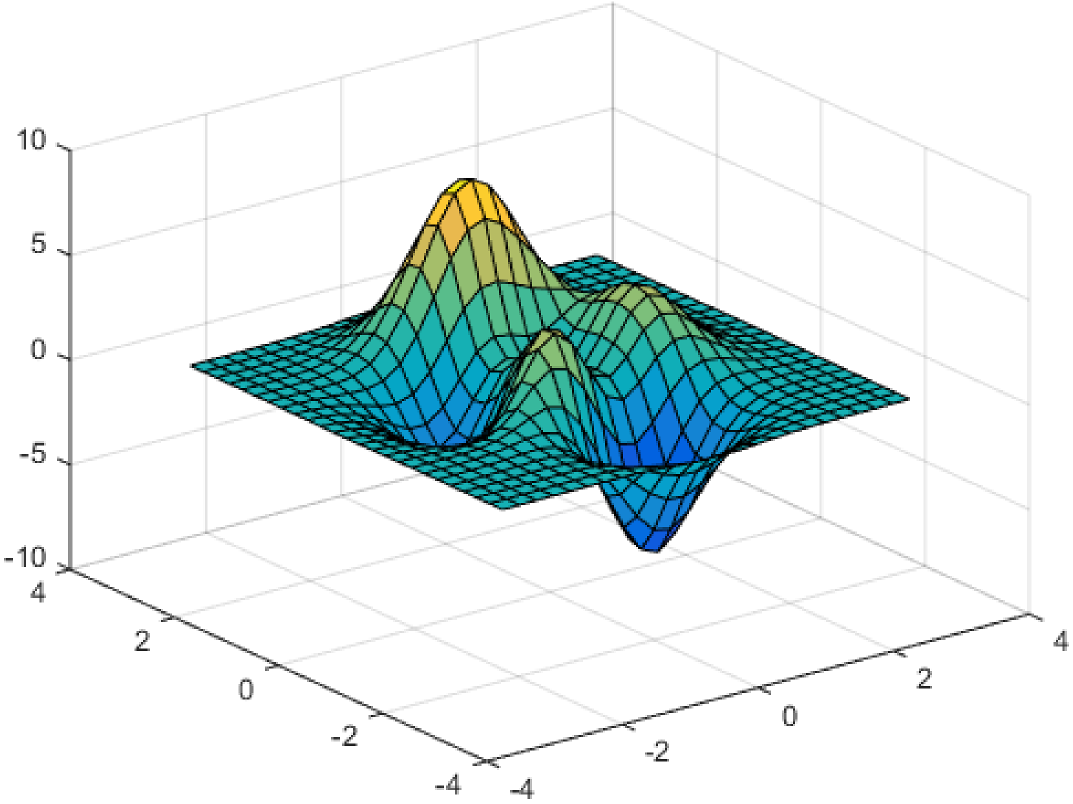
Data Columns | Sepal length | Sepal width | Petal length

VISUALIZING SPATIO-TEMPORAL DATA

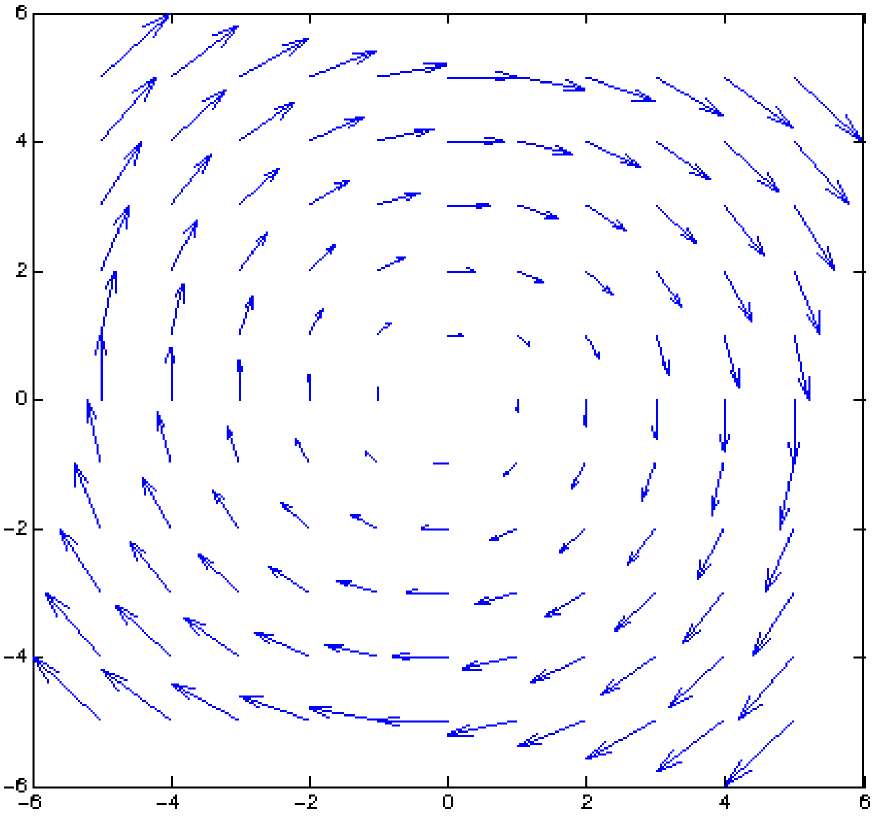
Contour plot



Surface Plot



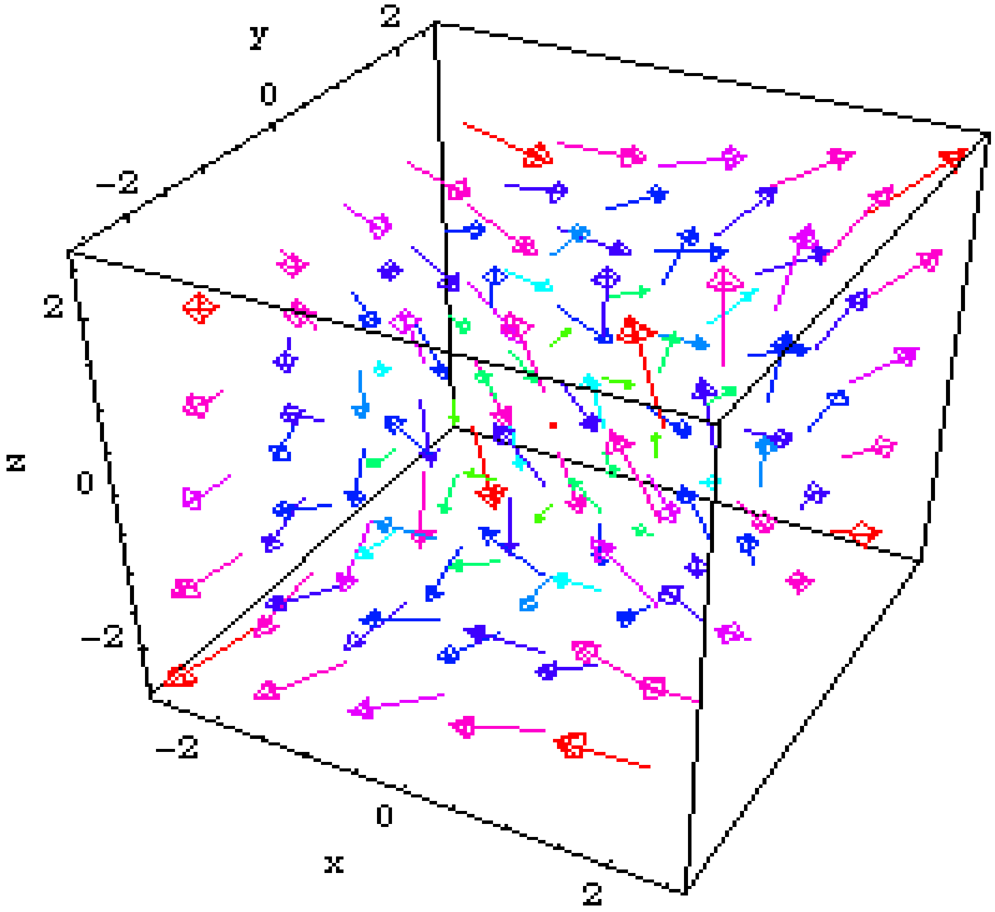
Vector field plots



2D

A vector is depicted at every point

Vector field plots

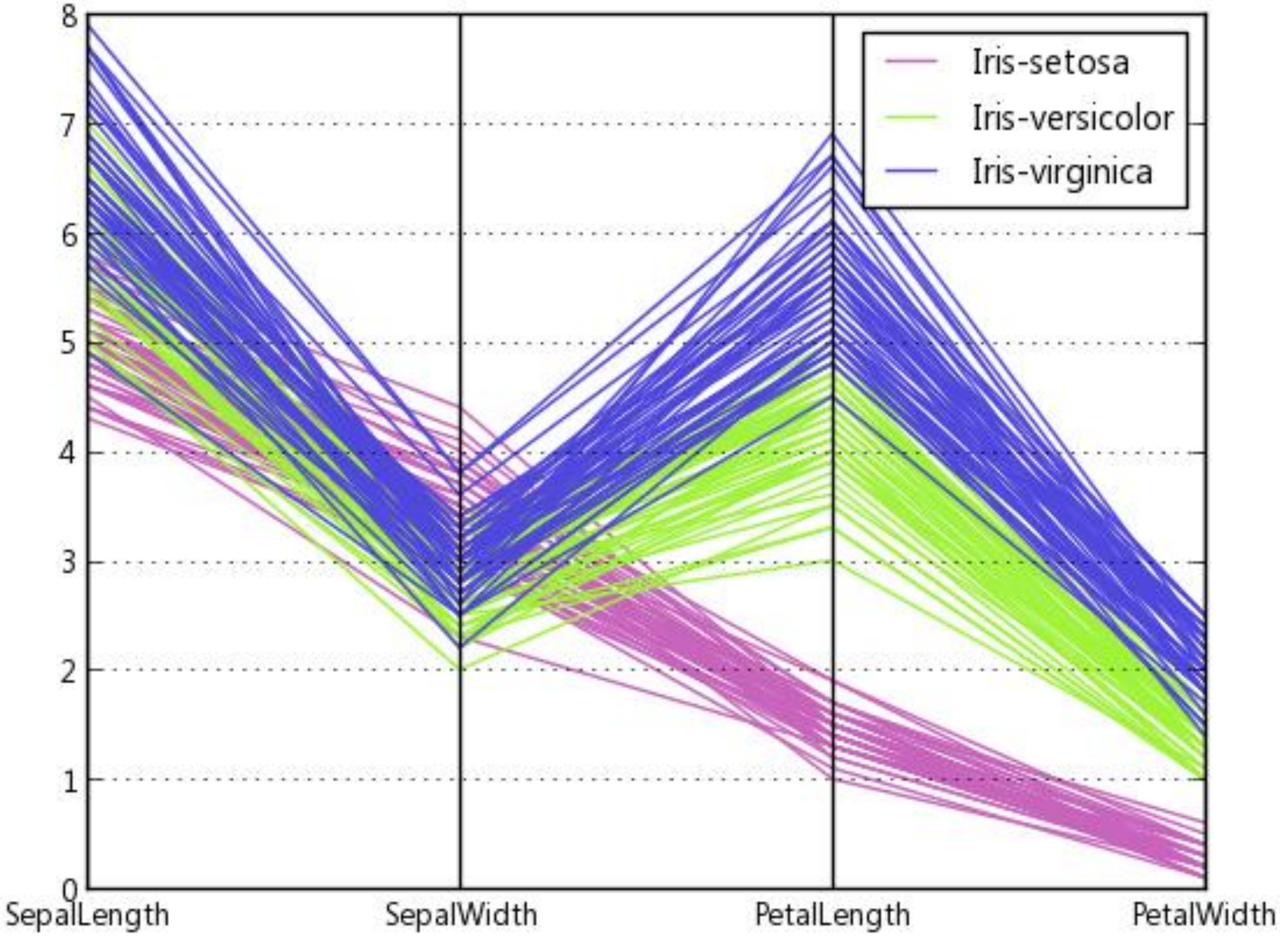


3D

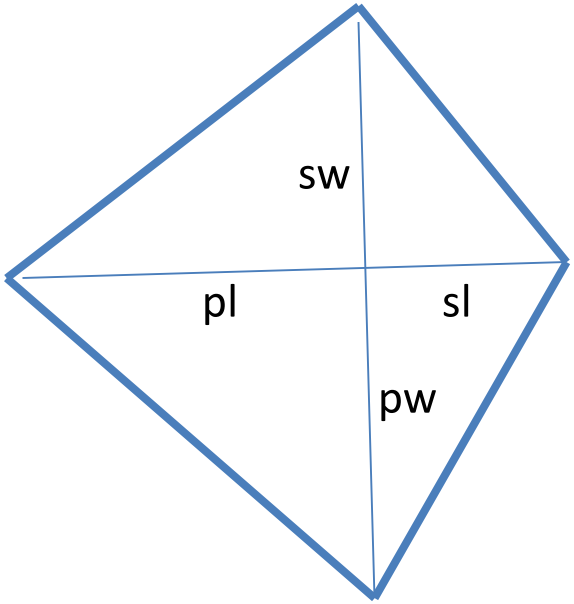
A vector is depicted at every point

VISUALIZING IN HIGH DIMENSIONS

Parallel Coordinates plot

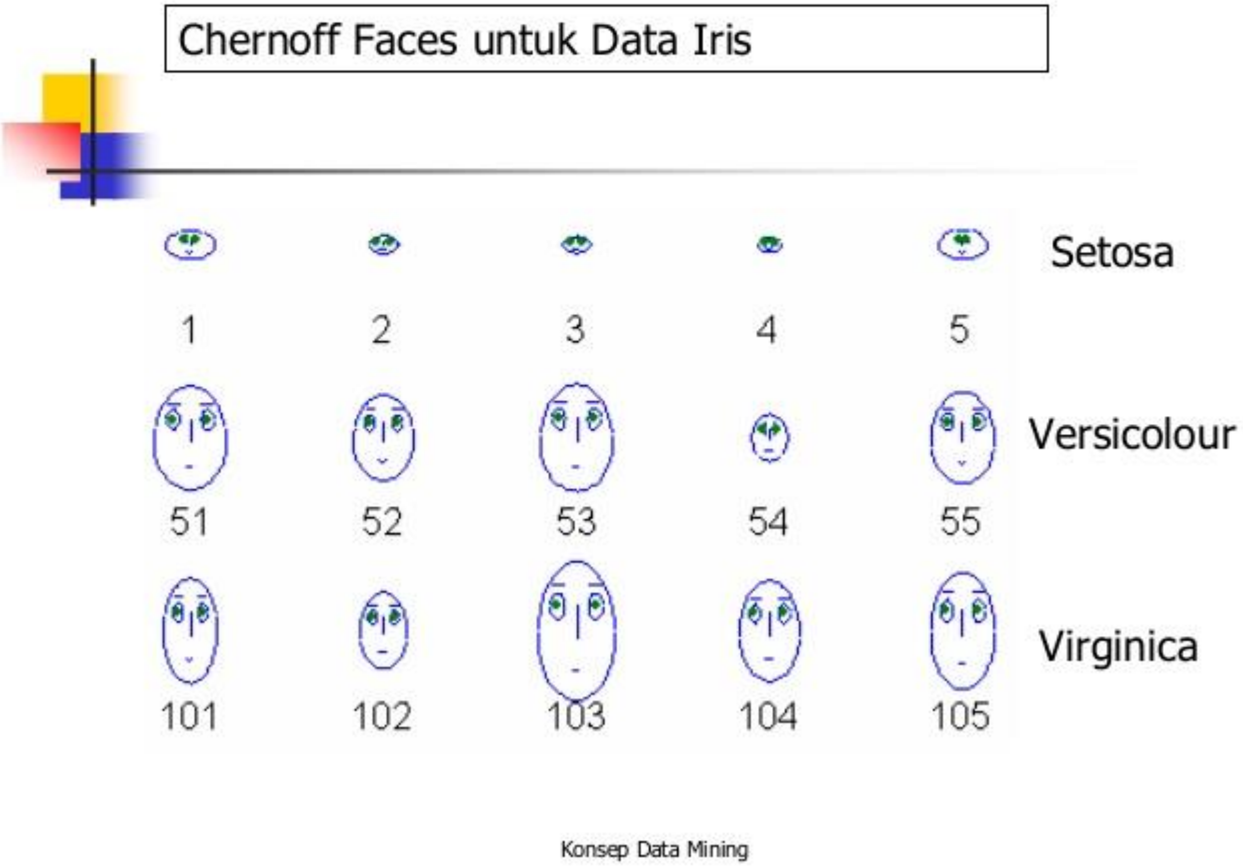


Star Coordinates



Sw – sepal width
Sl – sepal length
Pl – petal length
Pw – petal width

Chernoff faces

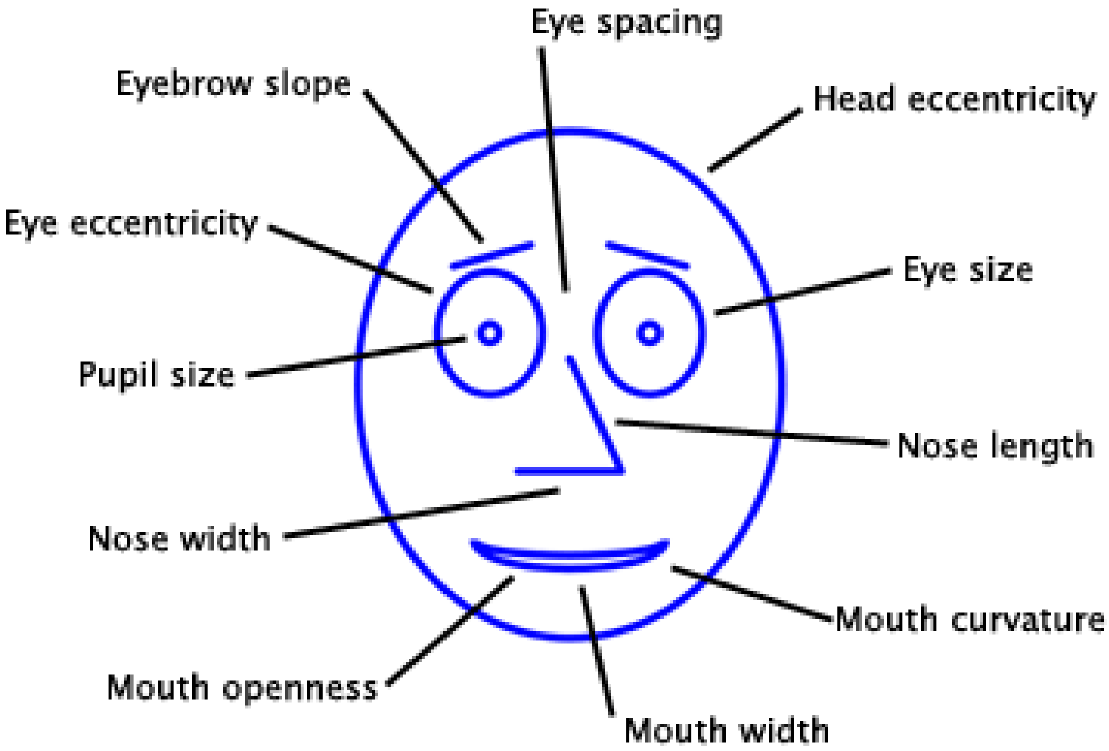


Chernoff Faces untuk Data Iris

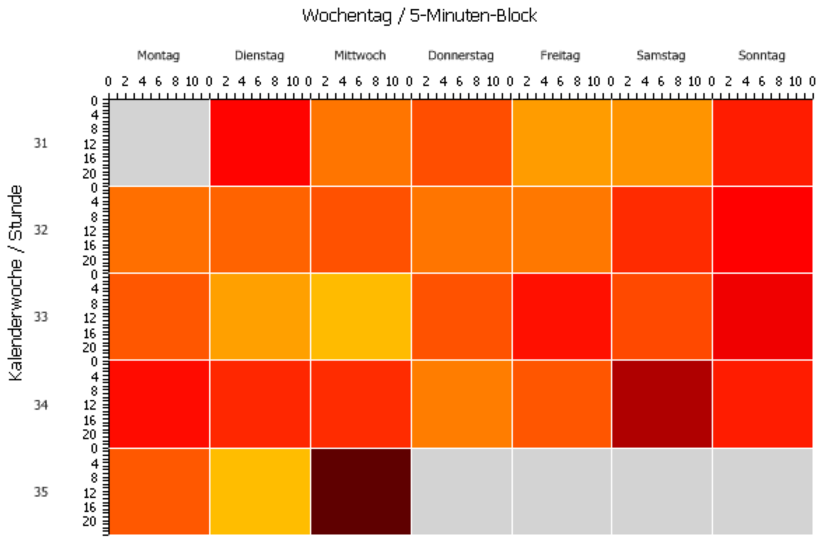
Konsep Data Mining

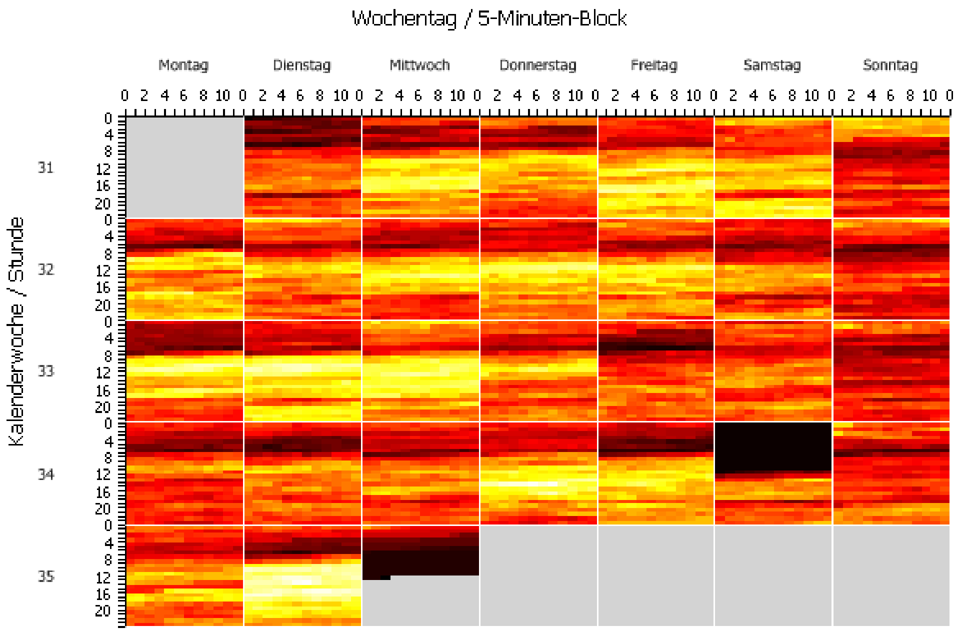
SL = size of face
SW = forehead/jaw ratio
PL = shape of forehead
PW = shape of jaw

A more intricate use of Chernoff faces



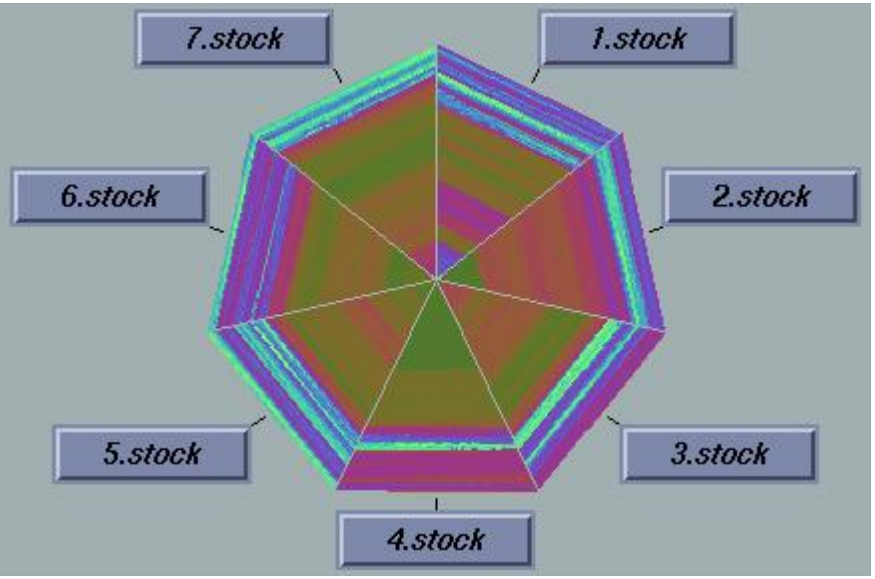
Multiscale visualization





(Shimabukuro et al 2004)

Circle segment display



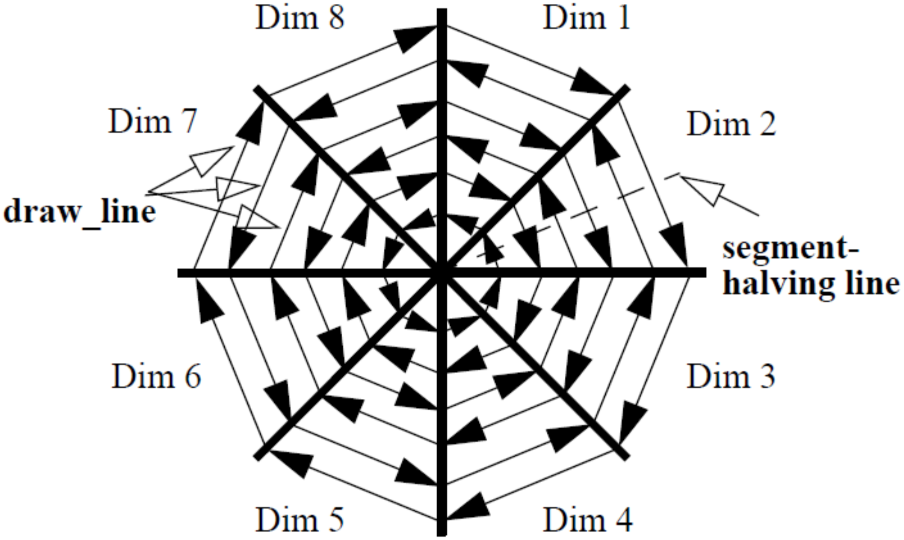
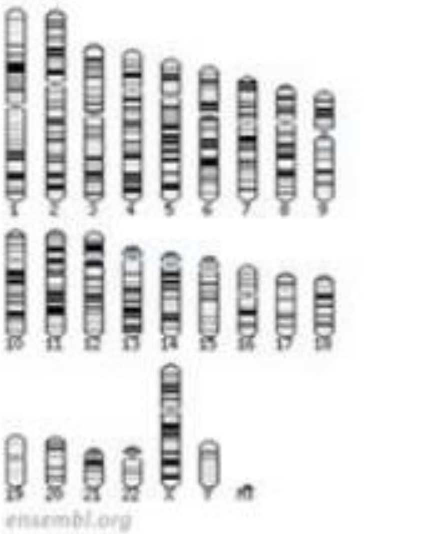
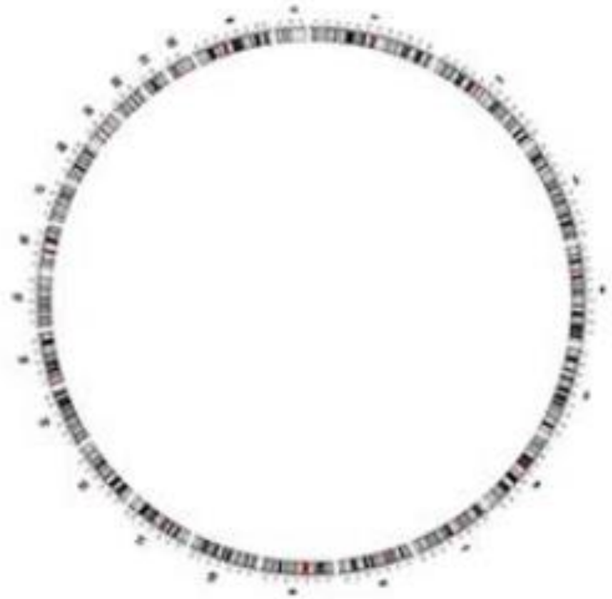


Figure 1: ‘Circle Segments’ Technique for 8-dimensional Data

CIRCOS

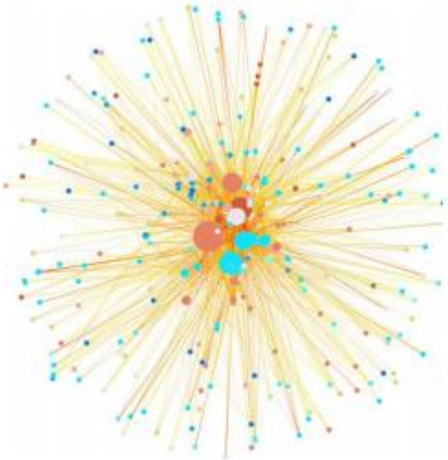


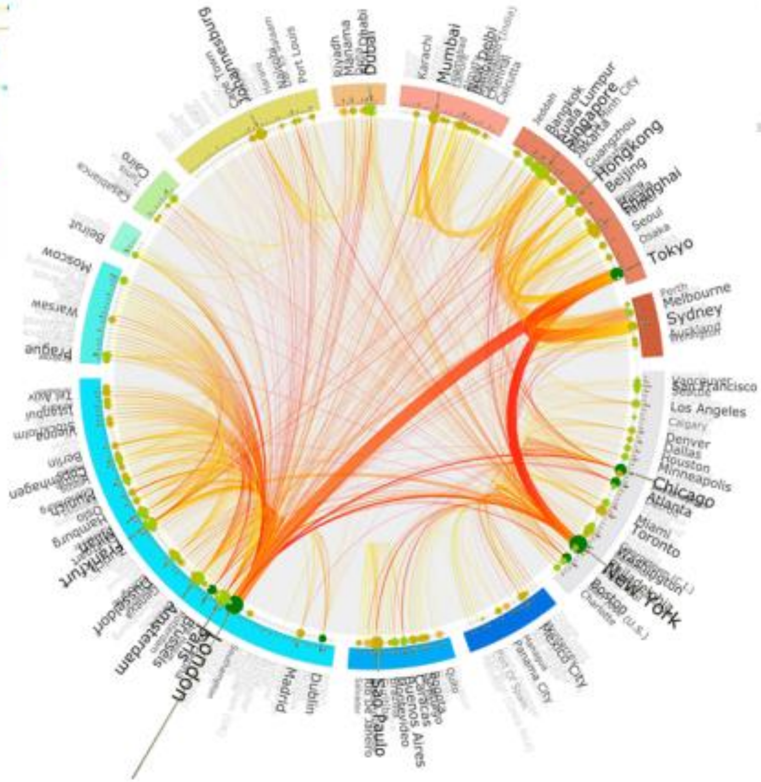
Classical ideogram layout



Circos ideogram layout

http://jura.wi.mit.edu/bio/education/hot_topics/Circos/Circos.pdf

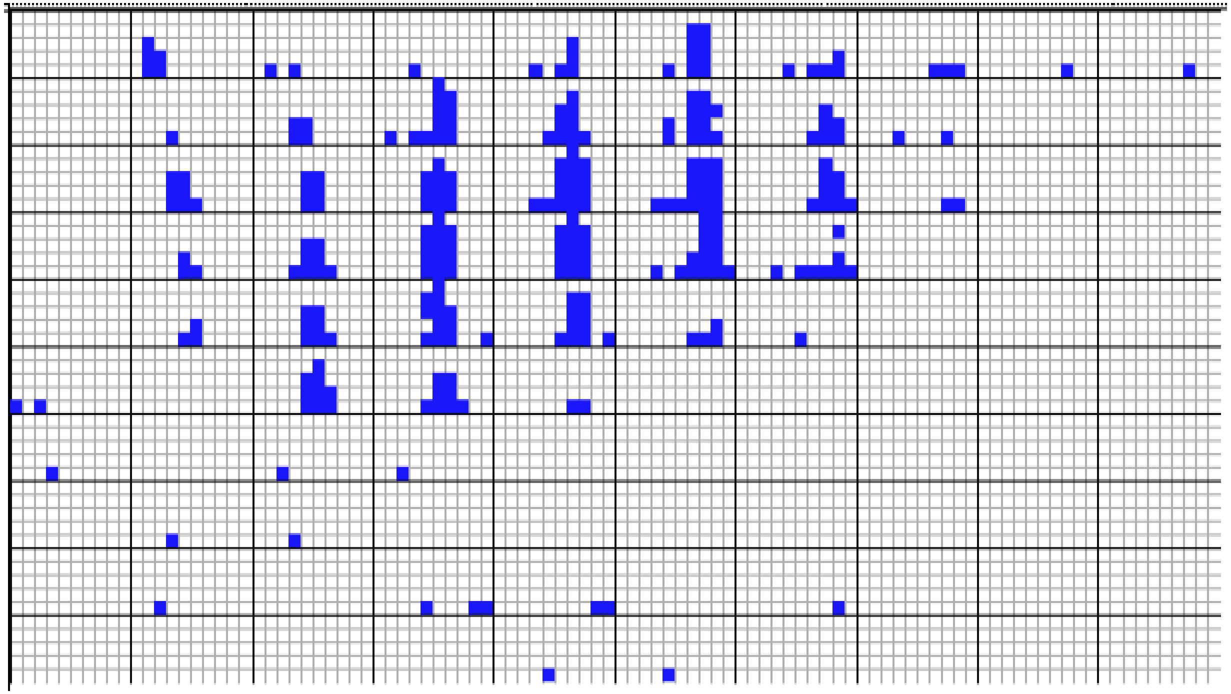




http://www.lboro.ac.uk/gawc/rb/rb421.html

Dimensional Stacking

Used by permission of M. Ward, Worcester Polytechnic Institute

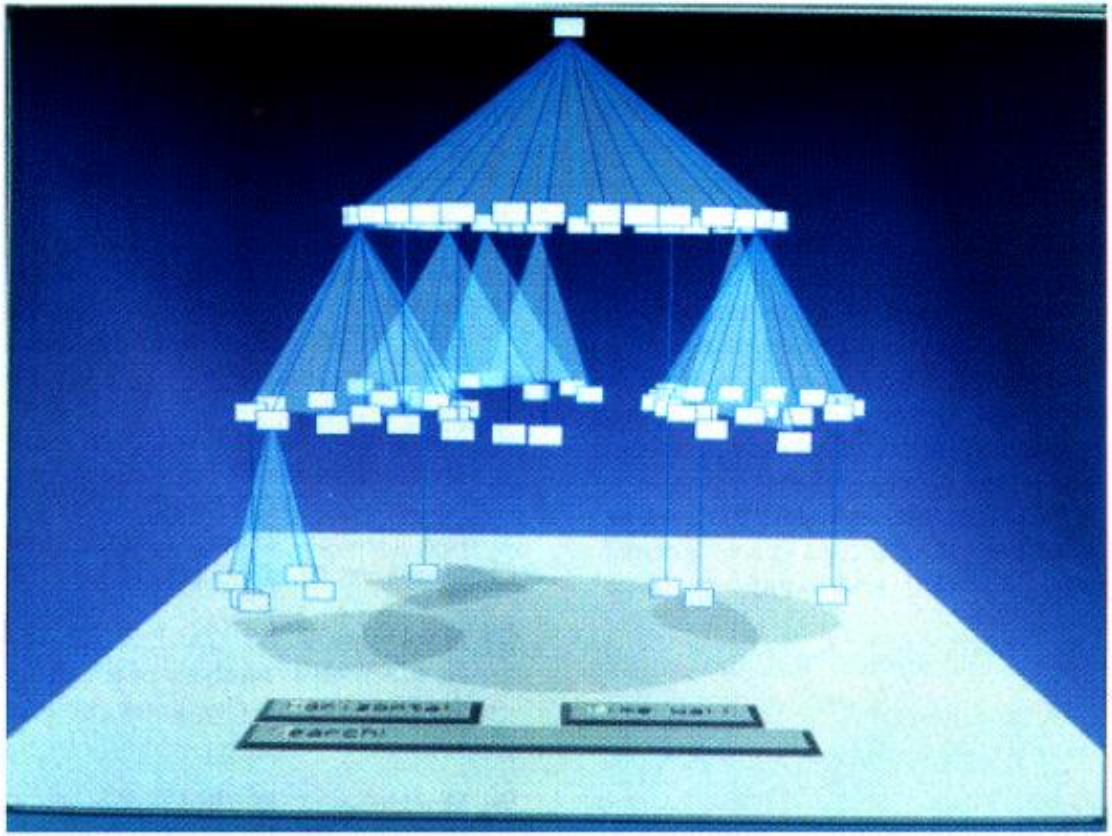


Visualization of oil mining data with longitude and latitude mapped to the outer x-, y-axes and ore grade and depth mapped to the inner x-, y-axes

Cone trees

- Display hierarchical data as cones
- Root node = apex; children = around the base
- Nodes are transparent so that you see the nodes in the background
- Cones lower in the hierarchy are progressively smaller
- If you click on a node, the node and the entire path from the root node are highlighted

A Cone Tree



Robertson Plate 1

Another cone tree!