Прогнозиране

Прогнозиране

- Данни
  - Числови данни
    - реални числа
  - Аналогови сигнали
    - дискретизиране
- Методи
- Регресия

Задача

- Да определи прогнозна стойност на един атрибут на данните (*target*), при известни стойности на другите атрибути (predictors)
- Примери
  - да определи очаквана стойност на заплатата на един специалист, в зависимост от трудовия му стаж
  - да определи адекватната продажна цена на жилищно имущество, в зависимост от големината му, близостта му до центъра и до училища.
- Модел
  - построен на базата на много предварителни наблюдения
- Построяване на модела (*training*)
  - алгоритми за откриване на функцията на зависимост между атрибутите и целта за всеки запис от обучаващата извадка
- Оценка на модела
  - изчисляване и анализ на грешките, породени от модела

Приложение

- Във всички области, в които обекти и процеси се измерват с числа
  - за
    - анализ на тенденции
    - планиране на бизнеса
    - маркетинг
    - научни изследвания
  - във
    - финанси и търговия – оценка на печалби, продажби
    - биология и медицина - оценка на ефект от алтернативни лечения
    - строителство и машиностроене – за моделиране на материали, конструкции
    - в екологията - поддържане на равновесие
- Когато една числова стойност е функция на един или повече фактори – числови величини

Методи за прогнозиране

Регресионен анализ

Регресионен анализ

- Метод за анализ на данни, при който се
  - изчислява една прогнозна стойност
  - може да изчисли интервал на прогнозната стойност, с определена степен на сигурност
- Обучение
  - supervised learning, *learning by examples*
  - търси се подходящо уравнение, което най-точно отразява данните
- Видове регресия
  - линейна
    - с една променлива
    - с много променливи
  - нелинейна

Регресионен анализ

- Метод от математическата статистика
- Цел: определяне на параметрите на функция, която обработва входните данни и прозвежда резултат, най-близък до наблюдавания
- Нотация
  - входни параметри - *predictors* (x_1 , x_2 , ..., x_n),
  - цел – непрекъсната величина (y)
  - търсена функция – $F(x,\theta)$
  - параметри на функцията - (θ_1 , θ_2 , ..., θ_n)
  - грешка (e)

$$y = F(x,\theta) + e$$

Линейна регресия

- Видове линейна регресия
  - с един атрибут (predictor)

$$y = \theta_2 x + \theta_1$$

  - с повече атрибути

    X – вектор от атрибути

    - не може да се визуализира в двумерно пространство
    - може да се представи чрез множество коефициенти на функцията на регресия, по един за всеки атрибут (predictor)

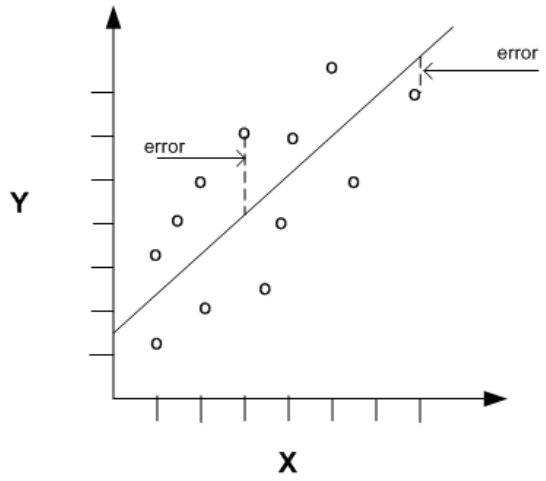
$$y = \theta 1 + \theta 2\ x1 + \theta 3\ x2 + \ldots\ldots \theta n\ xn\text{-}1 + e$$

Линейна регресия

- Пример за линейна регресия
  - апроксимация с права линия
- Коефициенти (параметри)
  - θ_1 – точката, в която x пресича координатата y ($x = 0$)
  - θ_2 - ъгълът на наклон

$$\theta_2 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

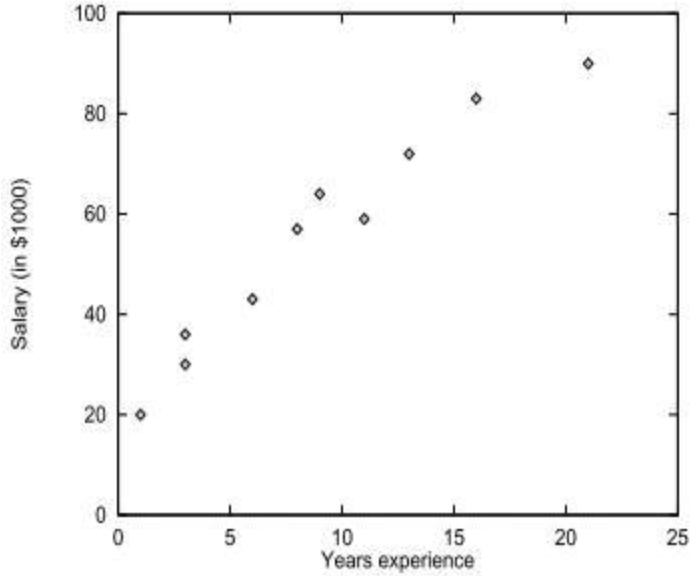
$$\theta_1 = \overline{y} - \theta_2 \overline{x}$$



Пример

- Да се прогнозира заплатата на човек с 10 години трудов стаж
- Обучаваща извадка данни

X *years experience*	Y *salary (in $1000)*
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83



Метод

- Изчисляване на средните стойности на x и y

  $\overline{x}= 9.1 \qquad \overline{y}= 55.4$

- Изчисляване на коефициентите

$$\theta_2 = \frac{(3-9.1)(30-55.4)+(8-9.1)(57-55.4)+\ldots+(16-9.1)(83-55.4)}{(3-9.1)^2+(8-9.1)^2+\ldots+(16-9.1)^2} = 3.5$$

$\theta_1 =$ 55.4 – 3.5 * 9.1 = 23.6

- Заместване в уравнението

  $y= 23.6 + 3.5\ x$

- Прогноза за x=10

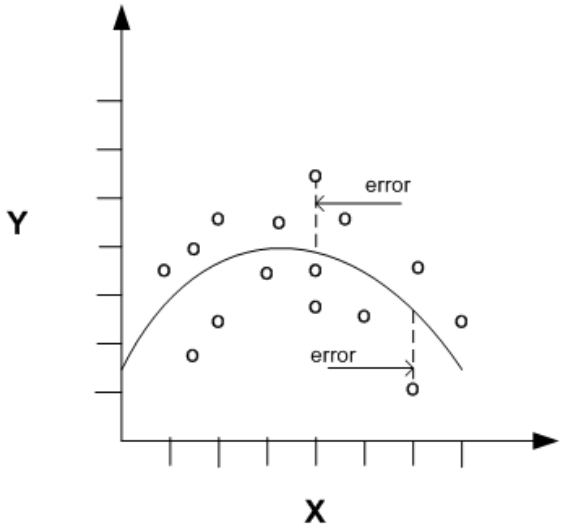
  y = 23.6 + 3.5 * 10 = 23.6 + 35 = 58.6 хил.

Нелинейна регресия

- В случаите, при които не е възможно апроксимиране с права линия
- Представяне чрез полином

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

- Възможно е заместване на кривата линия с множество прави линии



Нелинейна регресия

- Заместване

$X_1 = X$ $X_2 = X^2$ $X_3 = X^3$

- Преобразуване на уравнението в линейна форма

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

- Решаване на линейното уравнение по метода за най-малките квадрати

Оценка на грешката

- Метод
  - изчисляване на усреднена оценка на разстоянието от всяка точка до апросимиращата линия
- Мерки
  - Root Mean Squared Error (RMSE) – корен квадратен на сумата от квадратите на грешките на всички точки, разгледани поотделно

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

SQRT(AVG((predicted_value - actual_value) * (predicted_value - actual_value)))

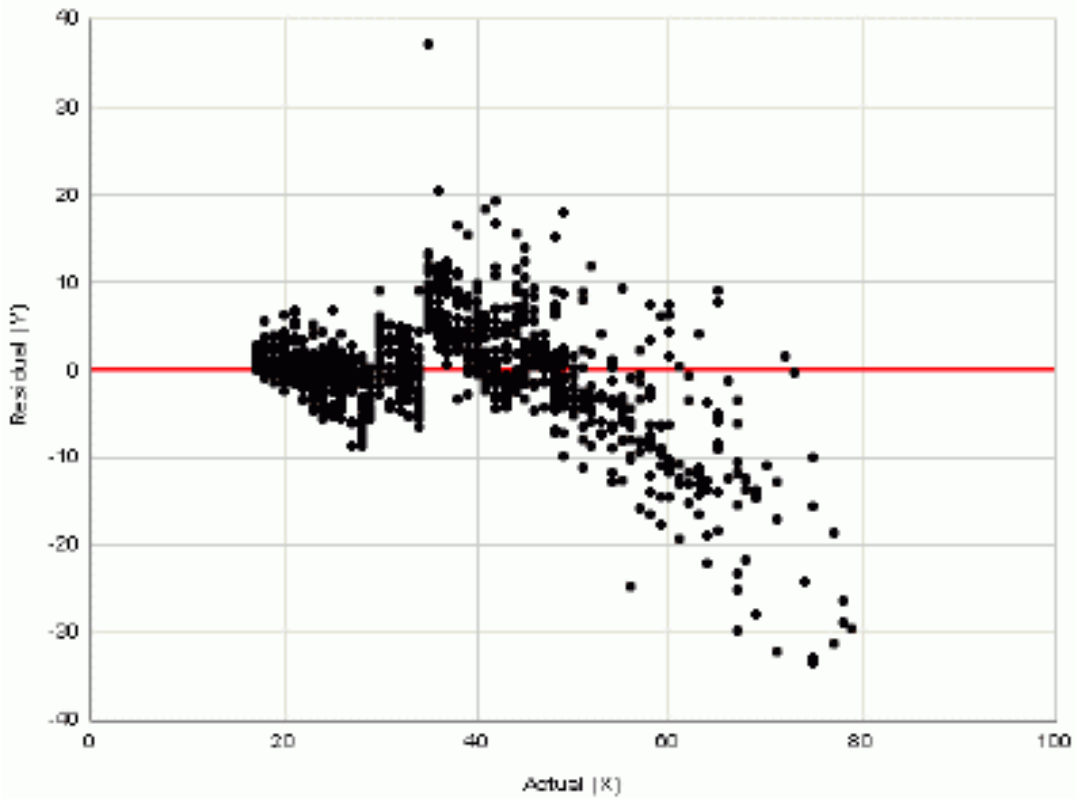
  - Mean Absolute Error (MAE) – абсолютна стойност на грешките

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

AVG(ABS(predicted_value - actual_value))

Residual Plot

- Диаграма на разликите между прогнозираните и действителните стойности



Support Vector Machines

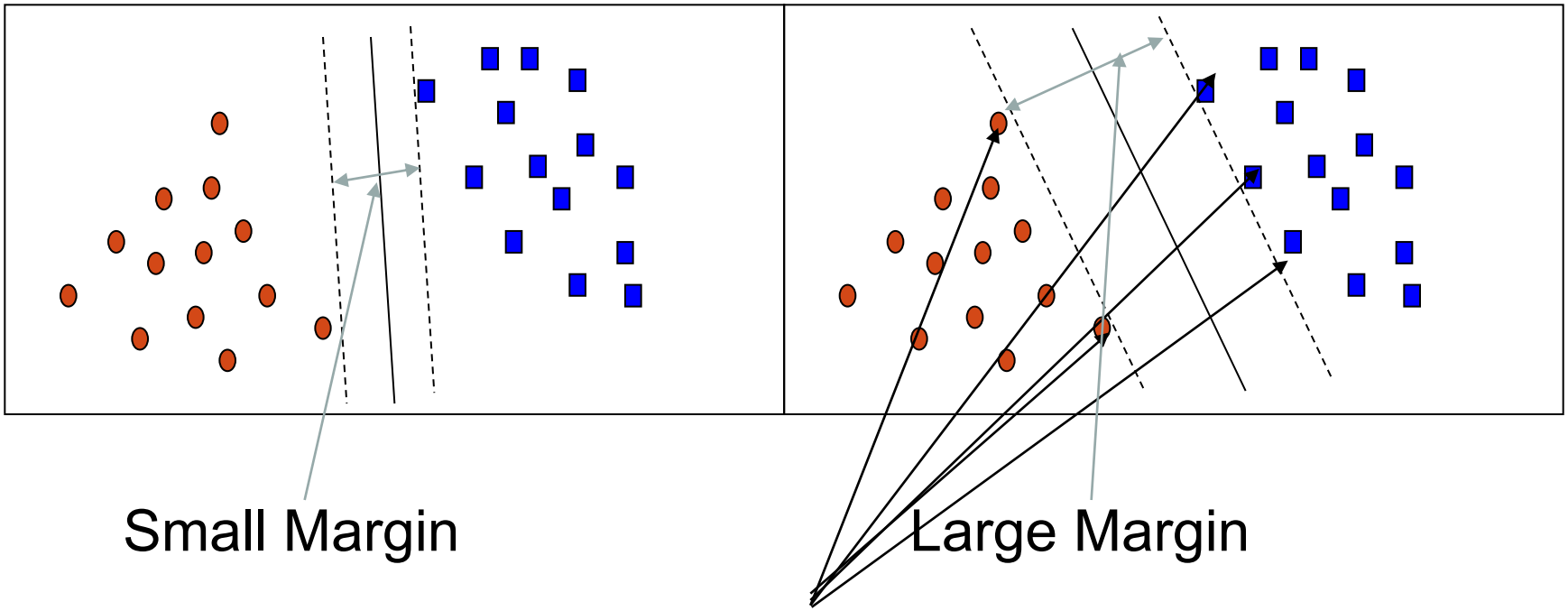
Support Vector Machines SVM

- Метод за прогнозиране и класификация
- Метод за линейна и нелинейна регресия
- Предимство: много висока точност
- Недостатък: бавен
- Прилага се за разпознаване на обекти
  - ръкопис
  - говор
  - образ

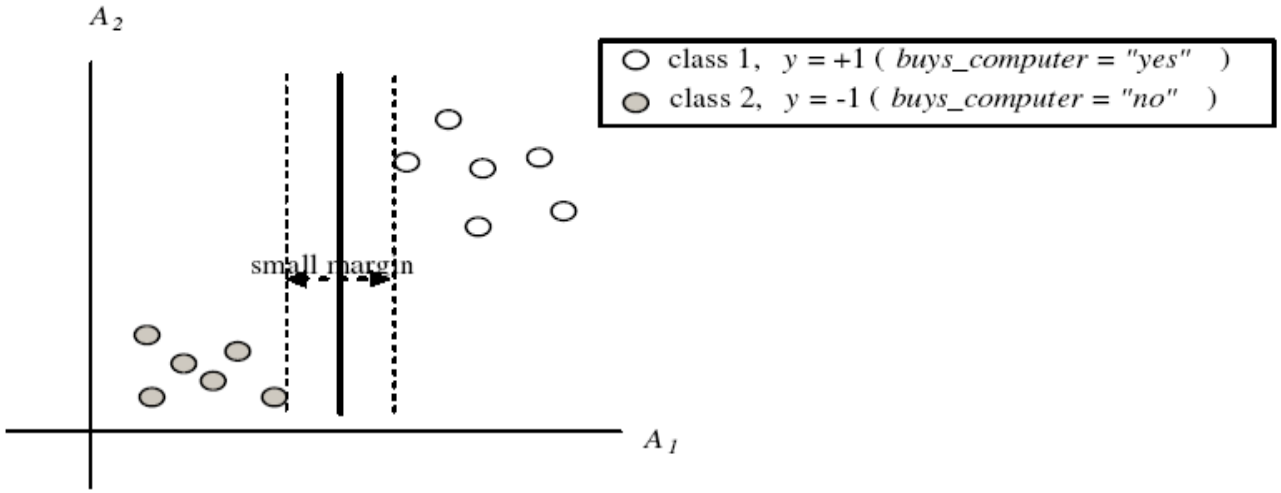
SVM

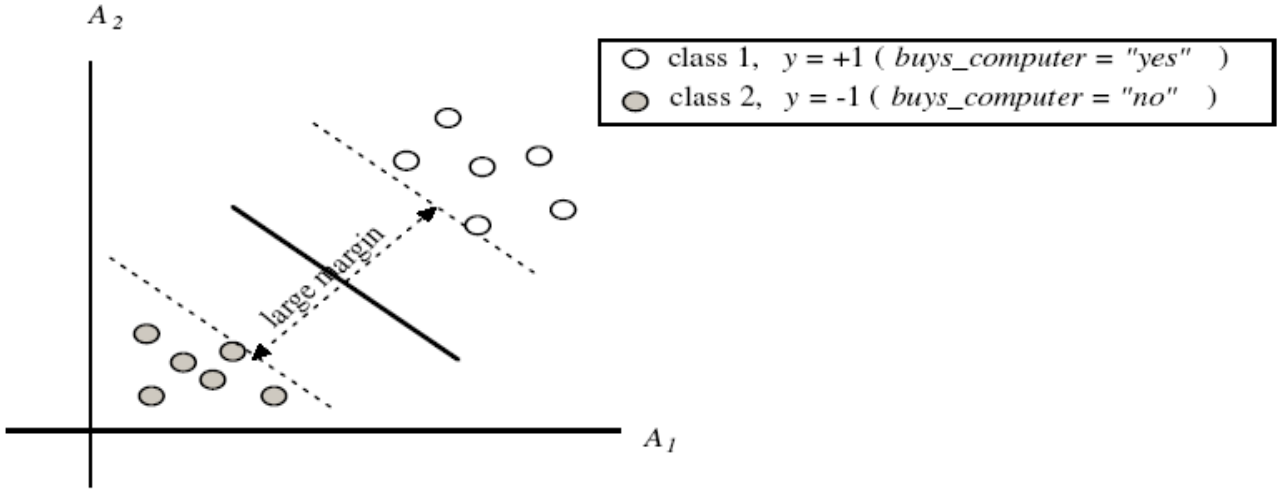
- Преобразува данните от обучаващата извадка в по-висока размерност, посредством нелинейни преобразования
- Търси **hyperplane** за оптимално разделяне на данните в два класа
- Два обекта
  - **support vectors** – важни записи от обучаващата извадка
  - **margins** – границите на областта, която разделя двата класа

SVM

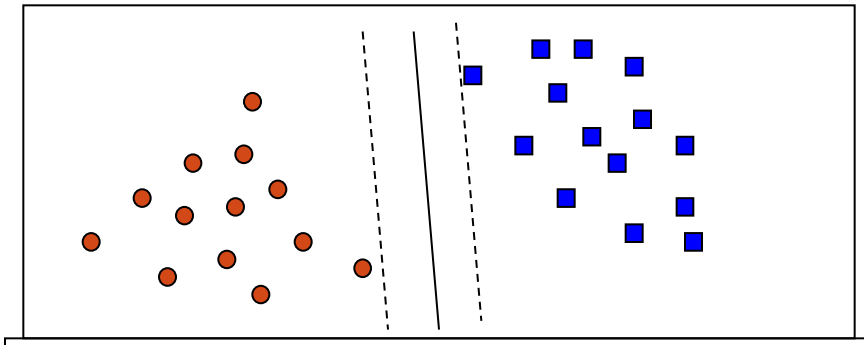


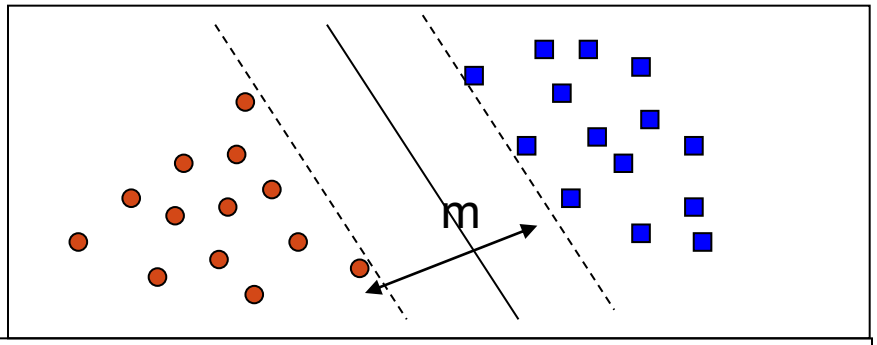
SVM


A_2
small margin
A_1
class 1, y = +1 (buys_computer = "yes")
class 2, y = -1 (buys_computer = "no")


A_2
large margin
A_1
class 1, y = +1 (buys_computer = "yes")
class 2, y = -1 (buys_computer = "no")

Нотация





D – данни,

($\mathbf{X}_1$, y_1), …, ($\mathbf{X}_{|D|}$, $y_{|D|}$) – множество от записи в обучаващата извадка

y_i – класове на принадлежност

hyperplane – пространство, което разделя класовете – линия, повърхнина или многомерен плот

support vectors – точките, които лежат точно на границите на хиперпространството

Между два класа съществуват безкраен брой разделящи пространства

Търси се най-доброто – което води до минимална грешка => най-широкото, т.е. **maximum marginal hyperplane** (MMH)

Линейно разделяне

- Разделящото пространство (линия) може да се представи

  $$\mathbf{W} \bullet \mathbf{X} + b = 0$$

  където $\mathbf{W}=\{w_1, w_2, \ldots, w_n\}$ е вектор на коефициенти, а b е скаларна величина (bias)

- За 2-D: $w_0 + w_1 x_1 + w_2 x_2 = 0$
- Границите на пространството (хиперповърхнината)

  $$H_1: w_0 + w_1 x_1 + w_2 x_2 \geq 1 \quad \text{за } y_i = +1, \text{ и}$$

  $$H_2: w_0 + w_1 x_1 + w_2 x_2 \leq -1 \text{ за } y_i = -1$$

  т.е. всеки запис над H_1 попада в първия клас, а тези под H_2 - във втория клас
- Записите, които попадат точно върху H_1 или H_2 и удовлетворяват неравенството $y_i (w_0 + w_1 x_1 + w_2 x_2) \geq 1$, за $\forall$ i са страните, които дефинират разстоянието m, т.е. **support vectors**

Анализ на потоци от данни

Времето е един от атрибутите

Потоци от данни

- Комплексни, неструктурирани или полуструктурирани данни
- Големи количества
- Динамично изменящи се
- Примери
  - мултимедия
  - web услуги
  - комуникации
  - сензорни мрежи
  - сателитни излъчвания
  - и др.

Подходи

- Sampling – периодично заснемане на текущо случайно състояние
- Sliding Windows – наблюдаване на различни под-интервали от данните
- Histograms
- Sketches – представяне на данните в различна грануларност и скали

Визуализация на резултатите

Голямо значение за потребителите

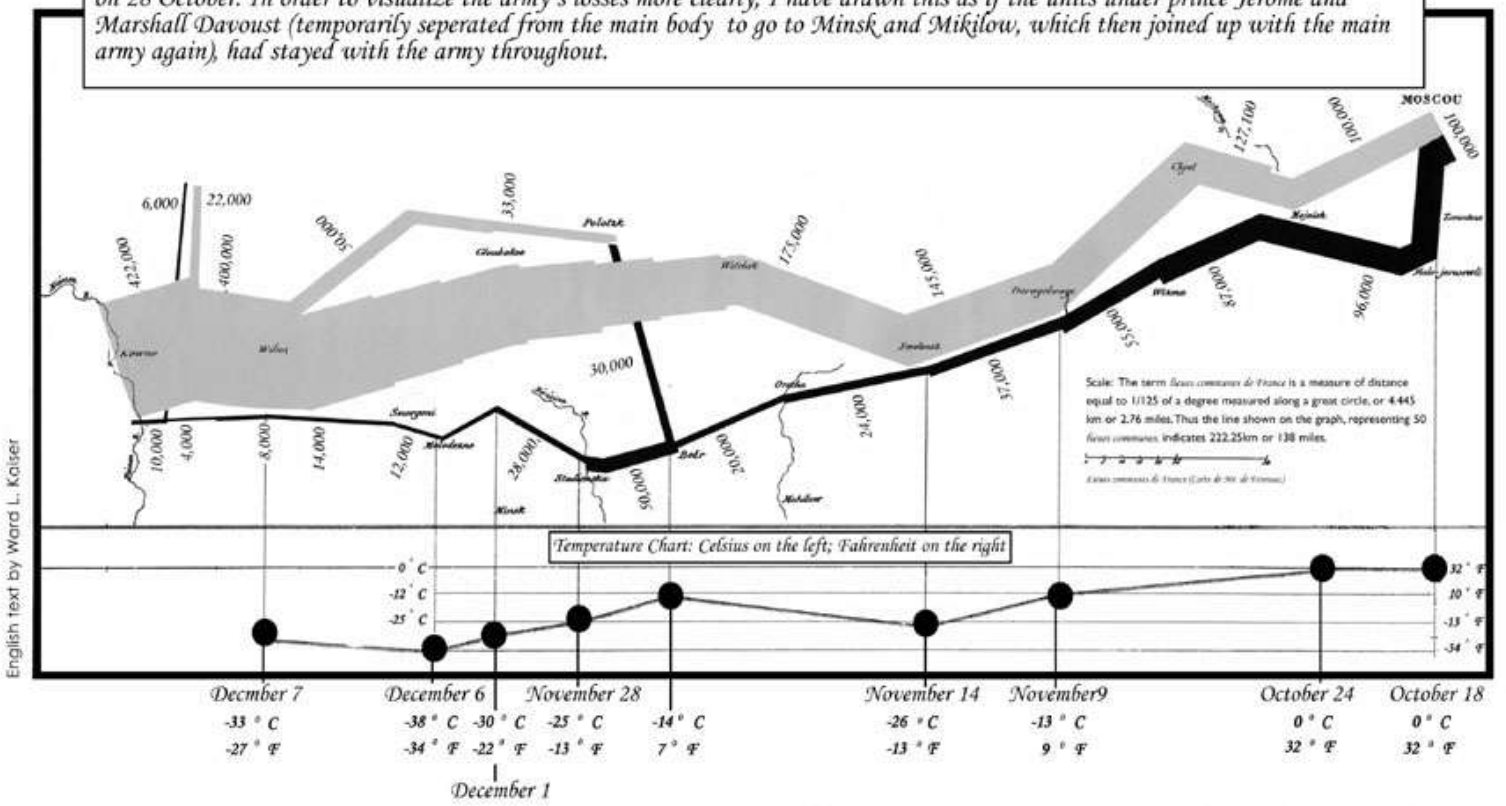
Визуализация

- Роля
  - за по-добро възприемане на резултатите за най-кратко време
  - подпомага итеративно изследване
  - изисква умения за разчитане
  - може да дава подвеждаща информация
- Средства
  - Карти
  - Диаграми
  - Многомерни изображения

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 seperated 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.

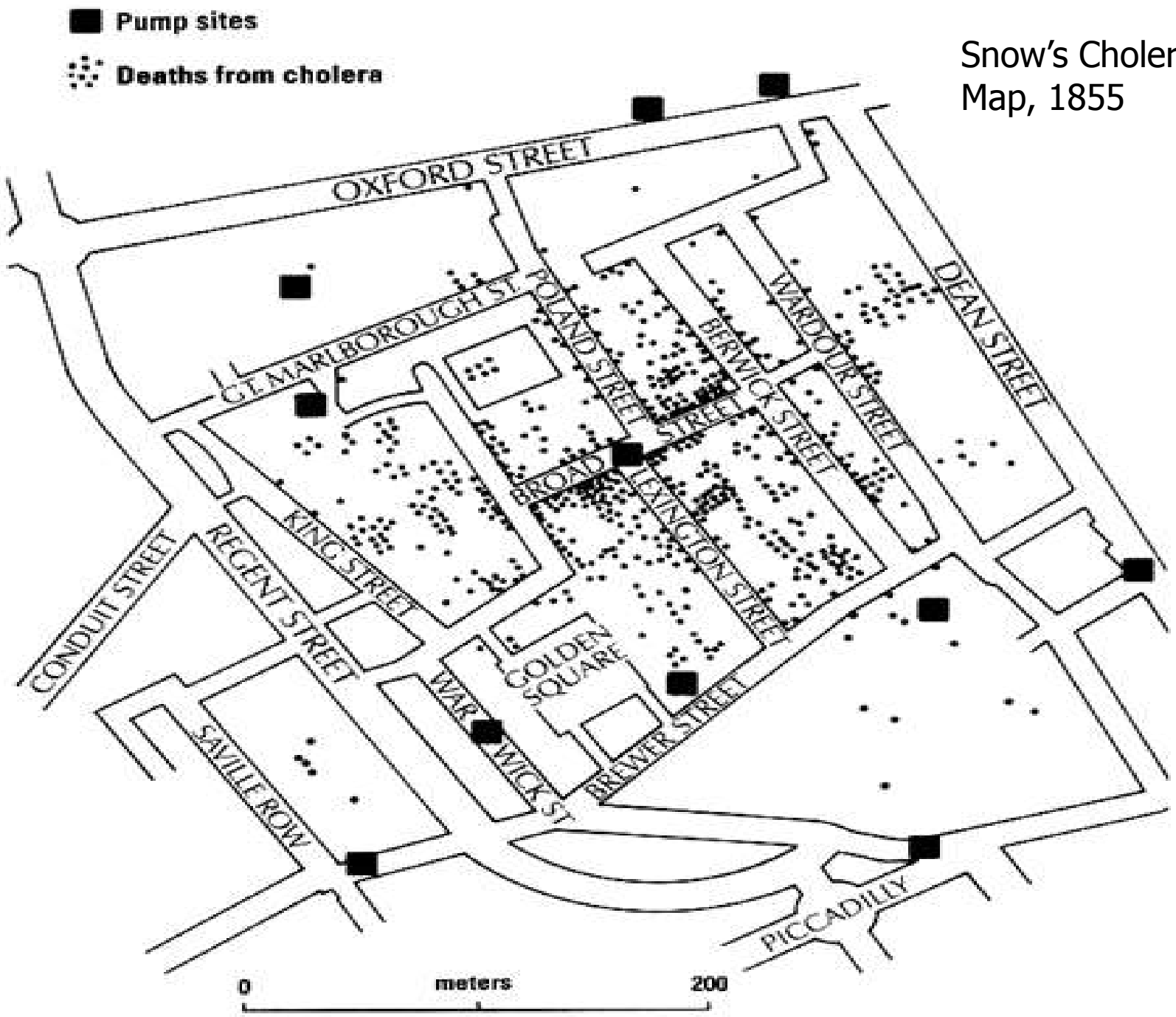


English text by Ward L. Kaiser

Editor s note: dates & temperatures are only referenced for the retreat from Moscow

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.***

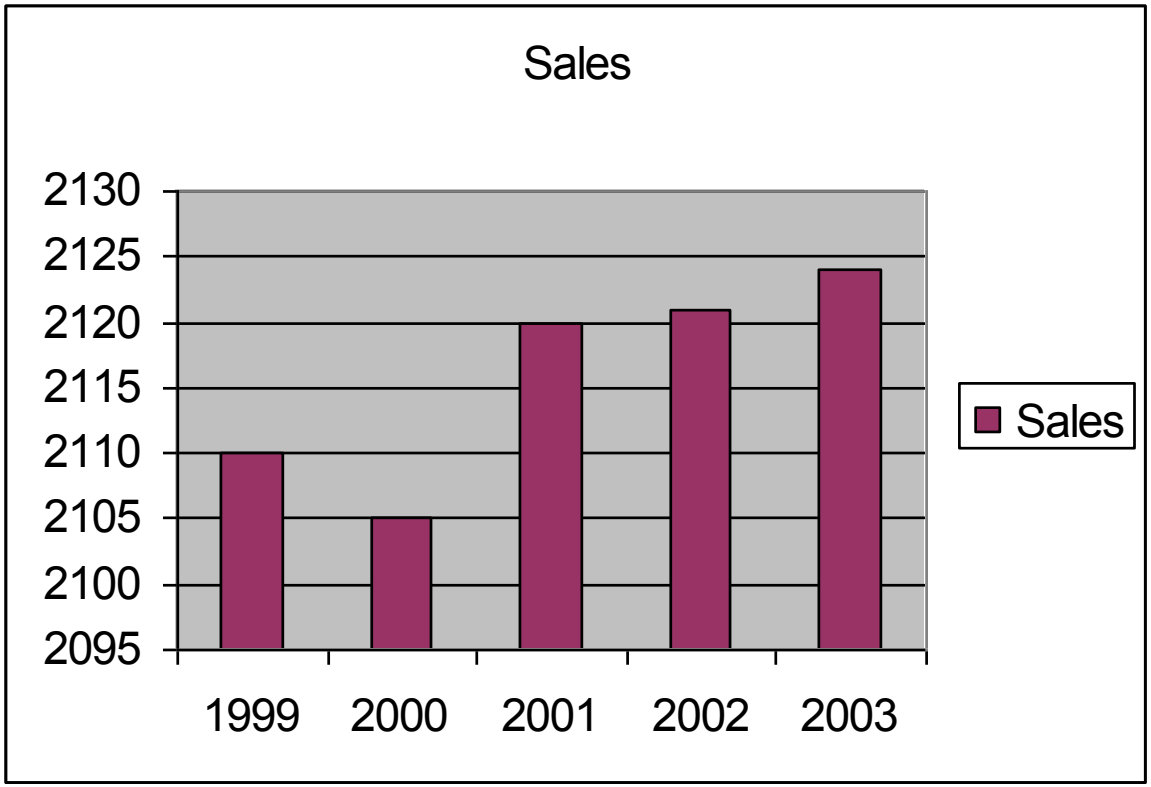
Snow's Cholera Map, 1855



Земята през нощта

Лоша диаграма

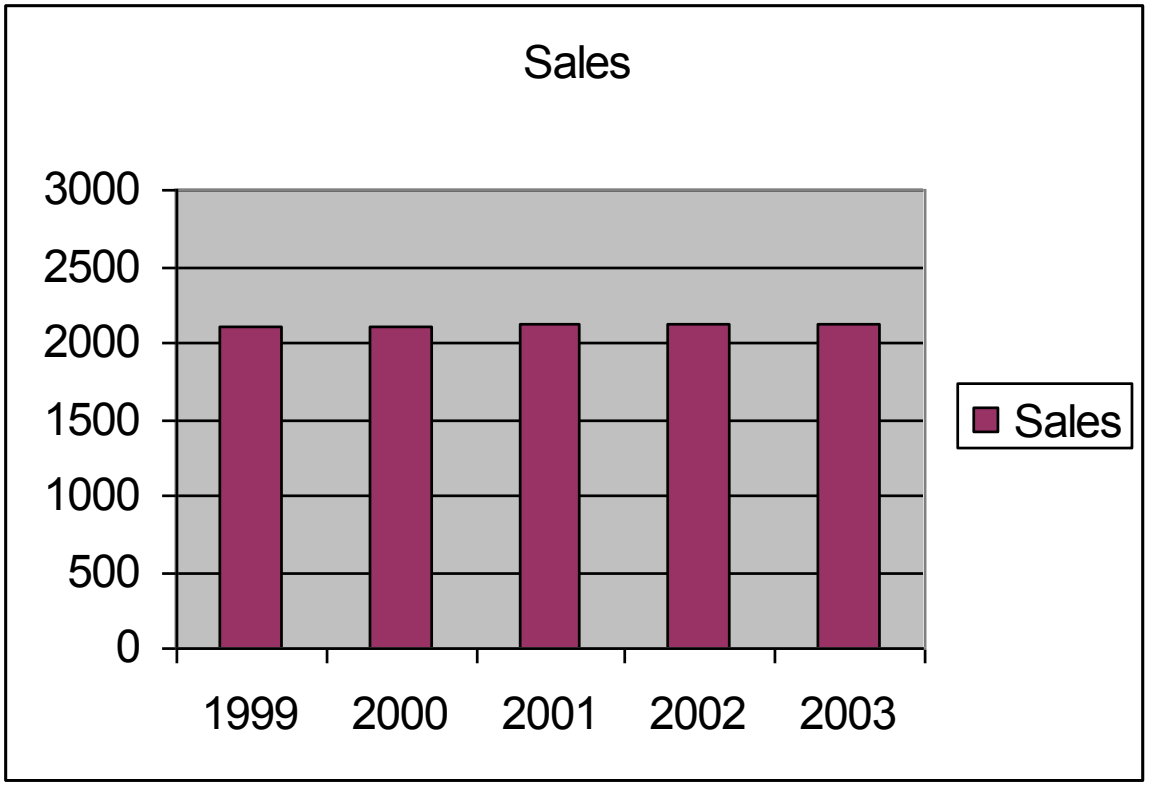
Year	Sales
1999	2110
2000	2105
2001	2120
2002	2121
2003	2124



Защо?

По-добра визуализация

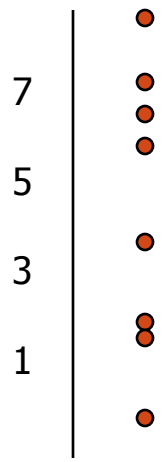
Year	Sales
1999	2110
2000	2105
2001	2120
2002	2121
2003	2124

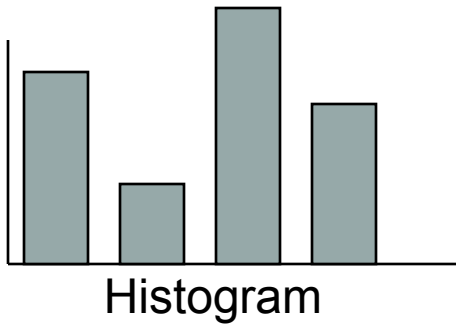


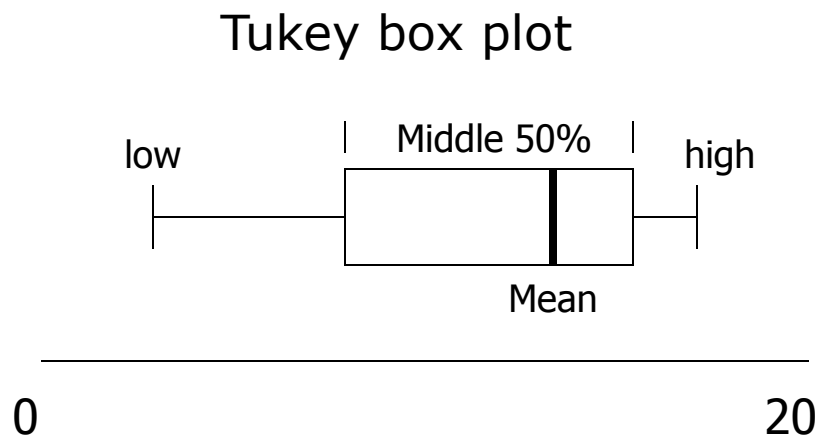
По-точна представа

1-D (Univariate) Data

- Представяне

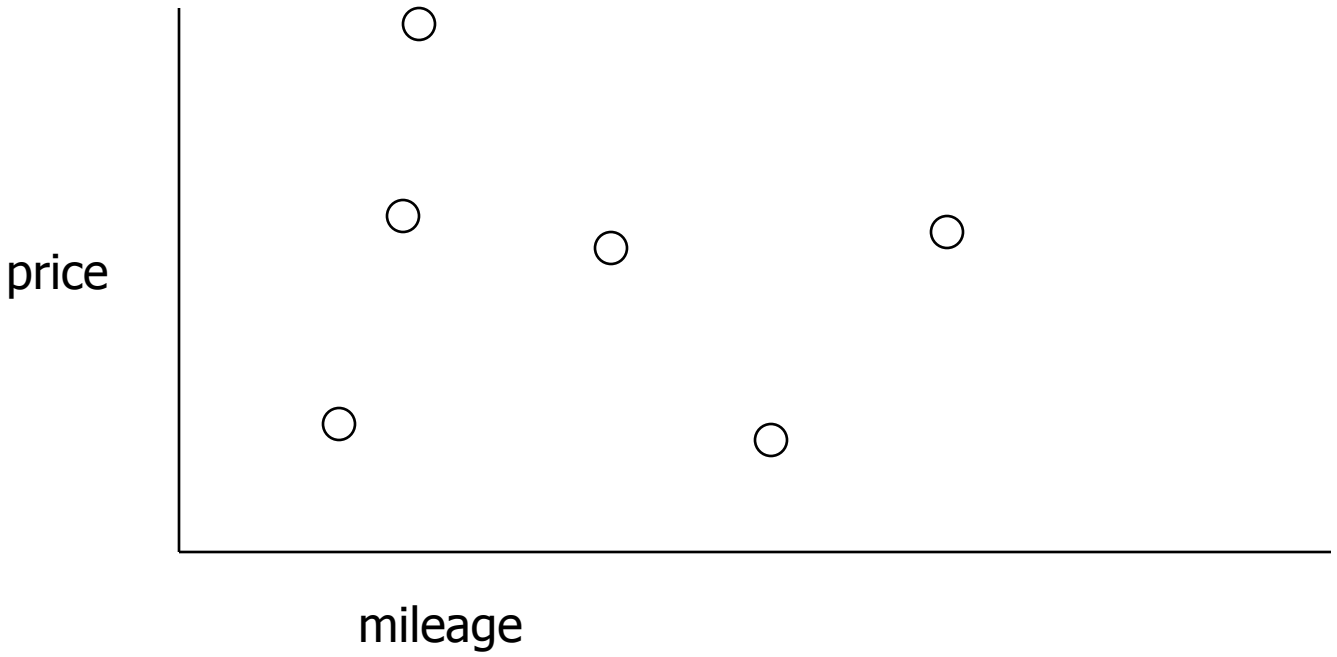




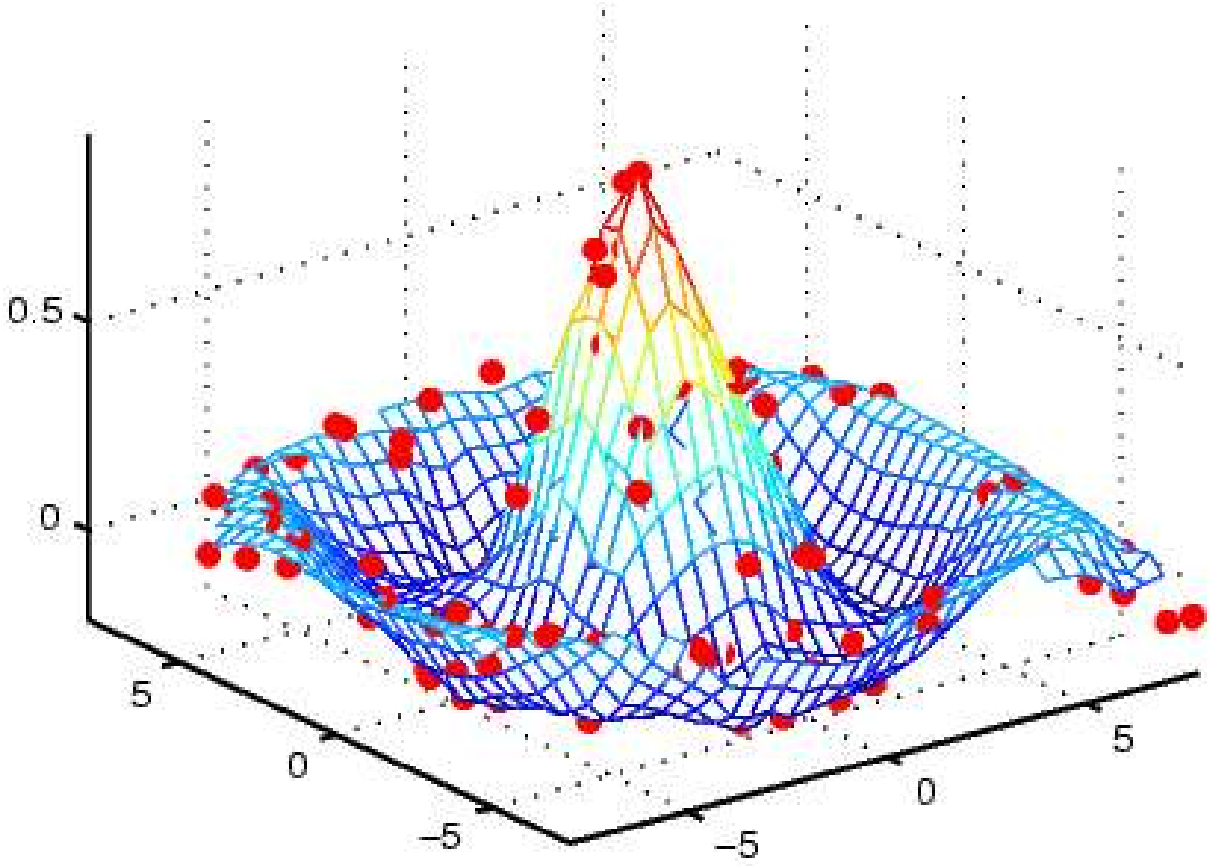


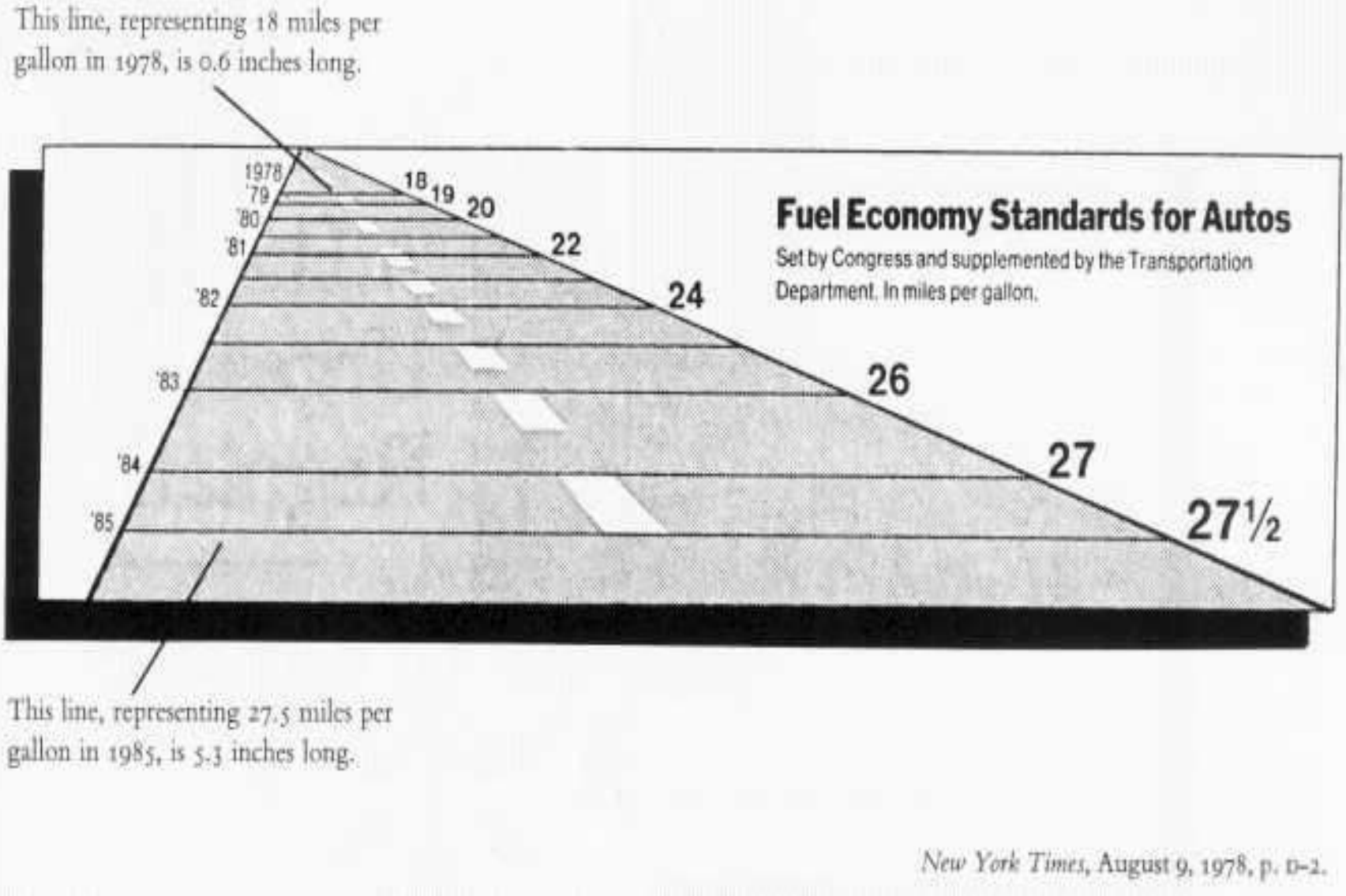
2-D (Bivariate) Data

- Scatter plot, …



3-D Data (projection)

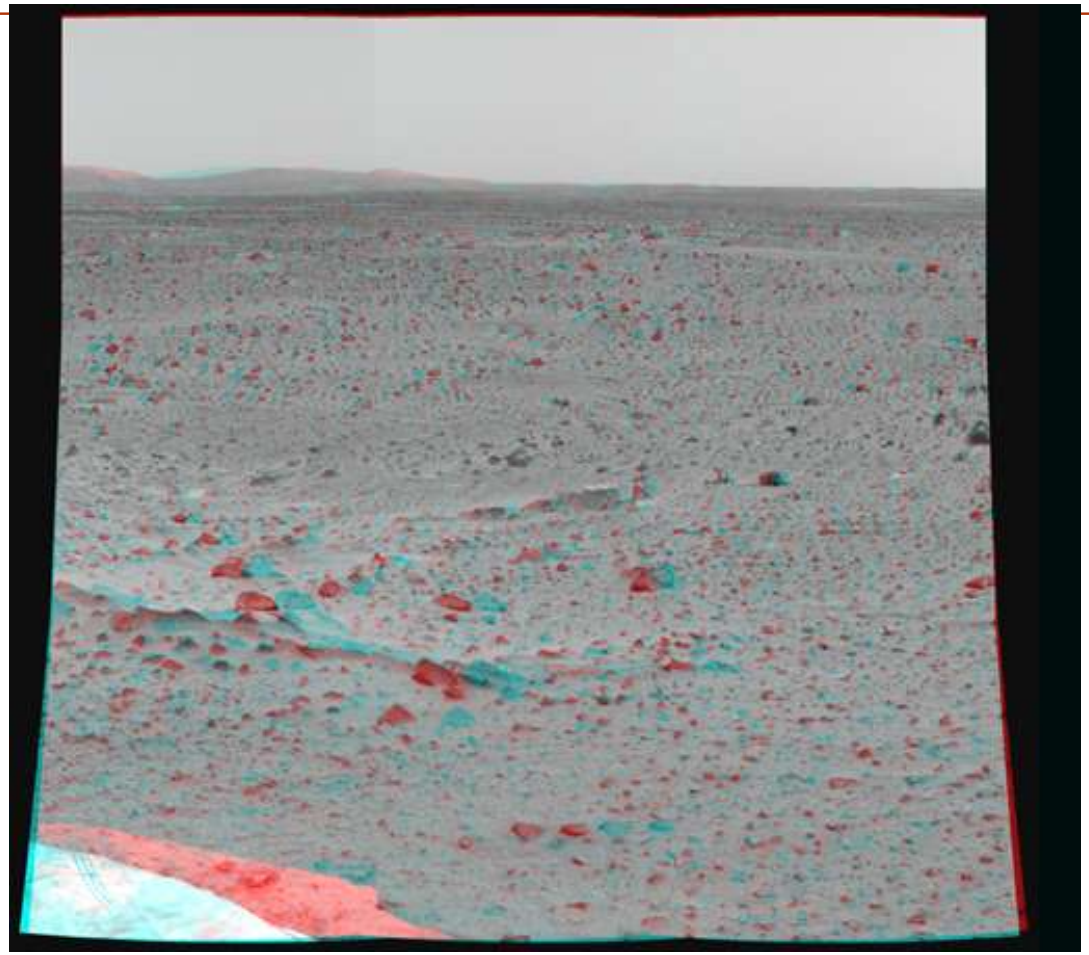




New York Times, August 9, 1978, p. D-2.

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

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

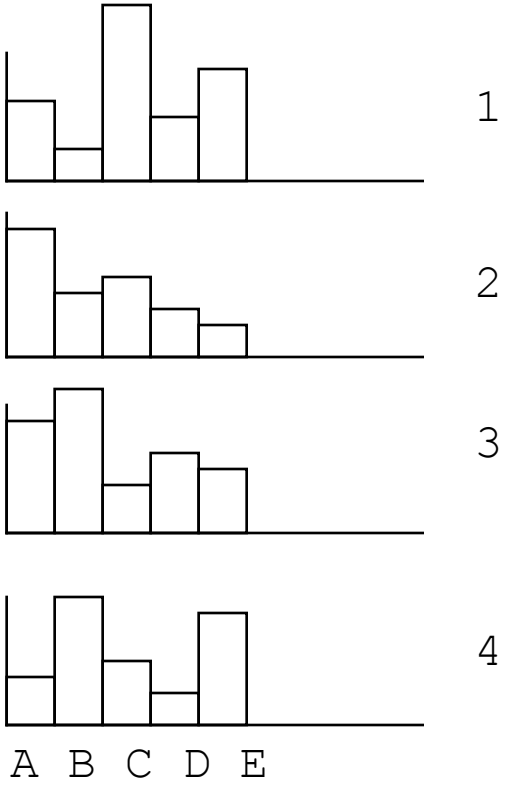


Taken by Mars Rover Spirit, Jan 2004

Multiple Views

Всяка променлива се илюстрира отделно

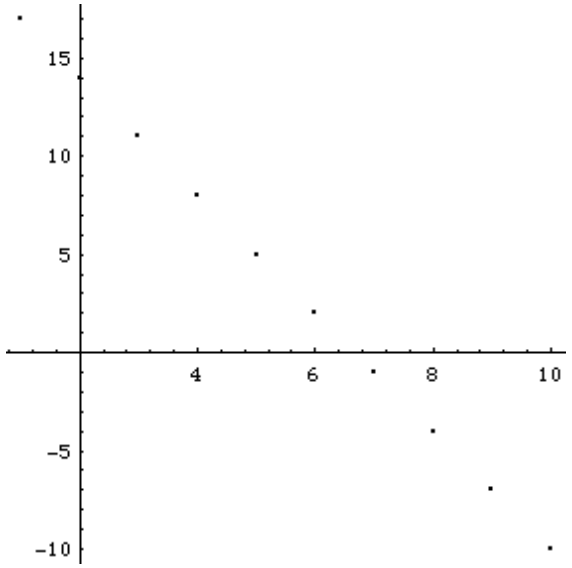
	A	B	C	D	E
1	4	1	8	3	5
2	6	3	4	2	1
3	5	7	2	4	3
4	2	6	3	1	5



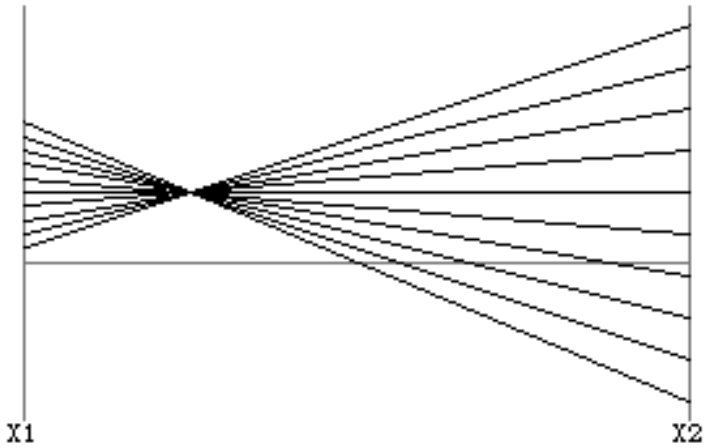
Недостатък: не показва отношенията между данните

Паралелни координати

- Наблюдаваните стойности – на хоризонталната линия
- По вертикалните линии – стойност на атрибут



В Декартови координати



В паралелни координати

Invented by Alfred Inselberg
while at IBM, 1985

Снимки и числа

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

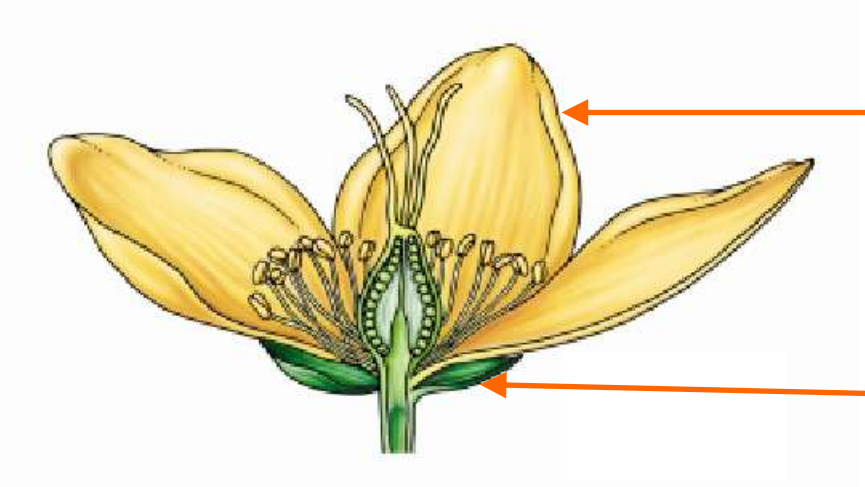
Iris versicolor

Iris virginica

не достатъюно информативно

Представяне в паралелни координати

Sepal Length



Атрибути на цвета

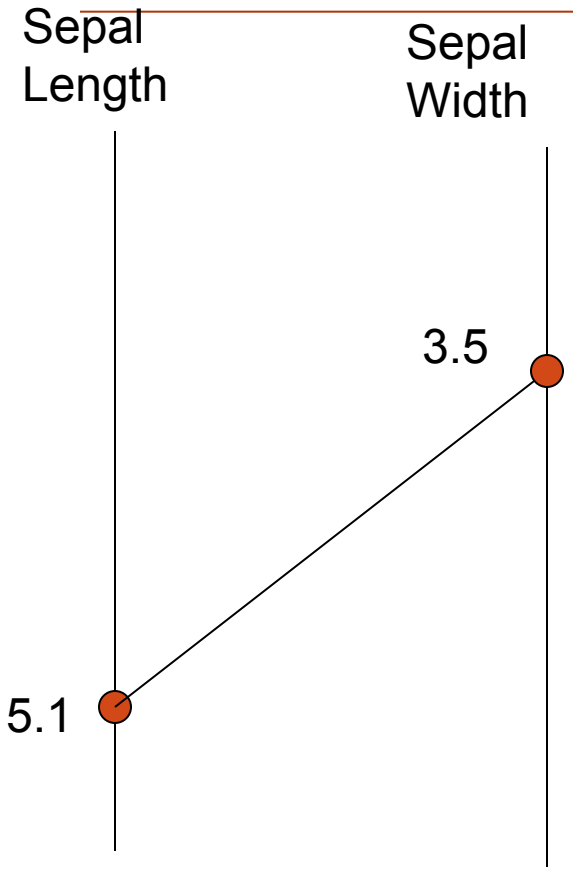
- Petal, a non-reproductive part of the flower

- Sepal, a non-reproductive part of the flower

5.1 Всеки атрибут се представя чрез линия, всяка стойност – като точка върху линията

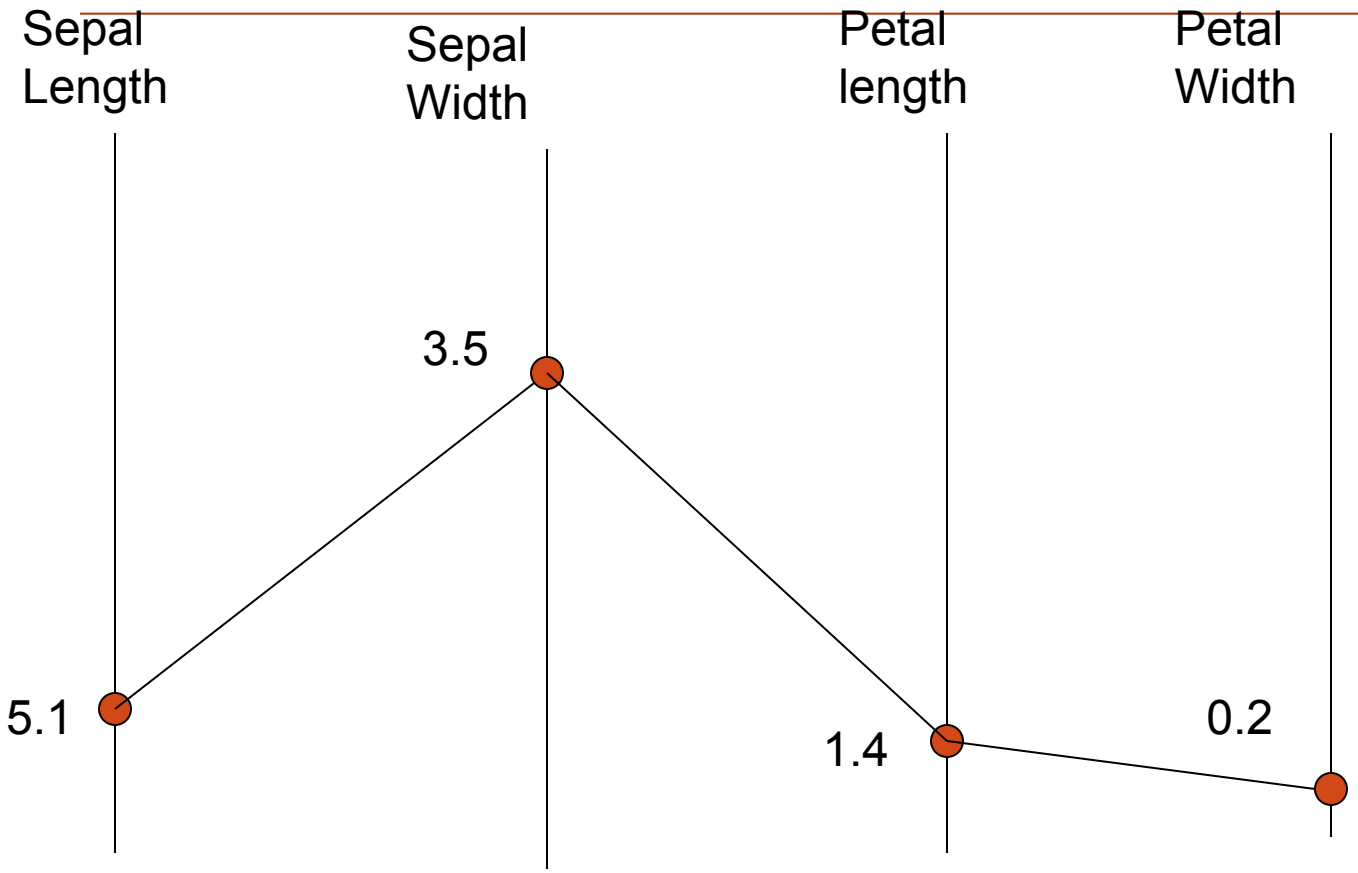
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Parallel Coordinates: 2 D



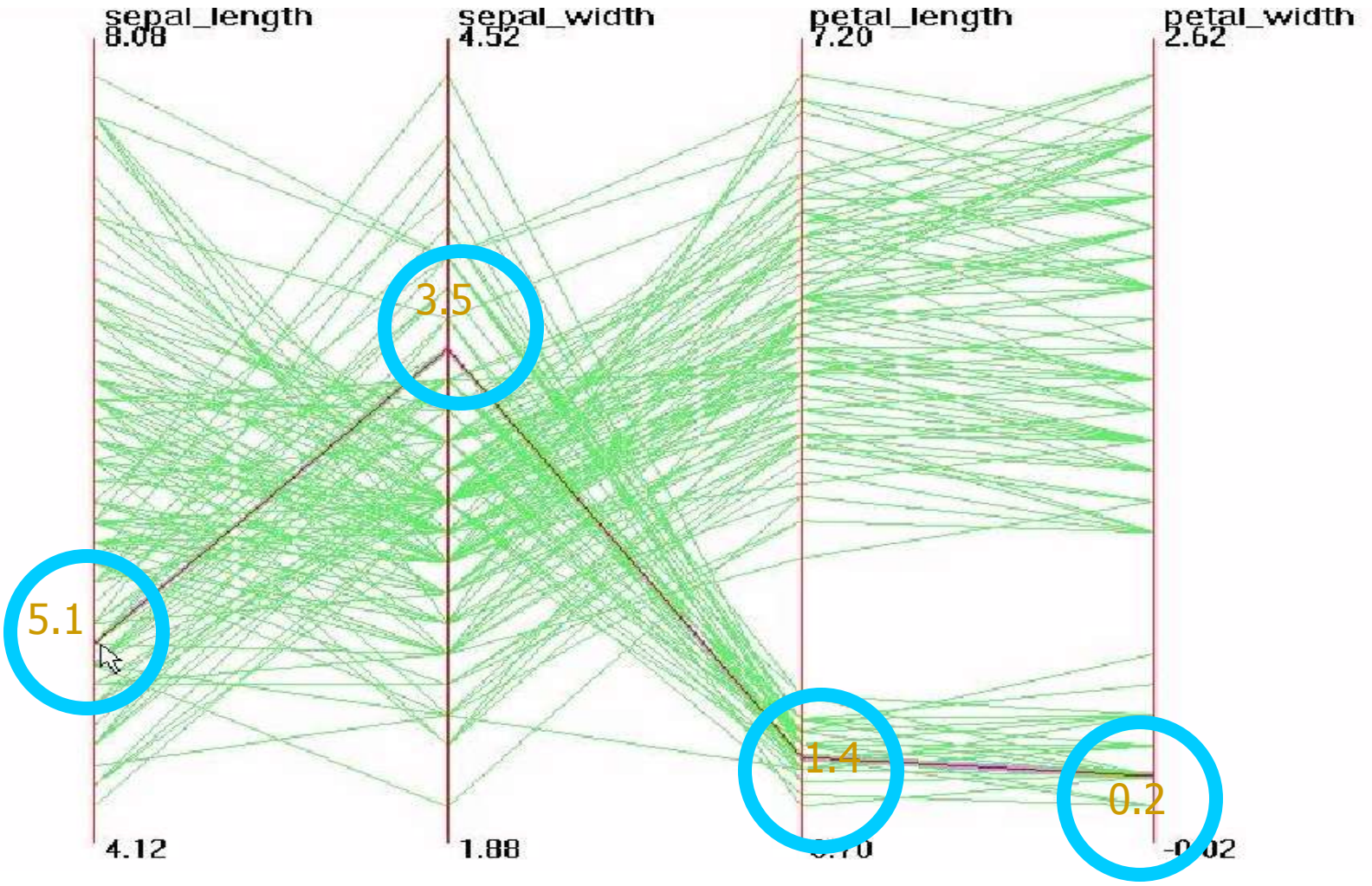
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Паралелни координати: 4 D



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

Паралелна визуализация

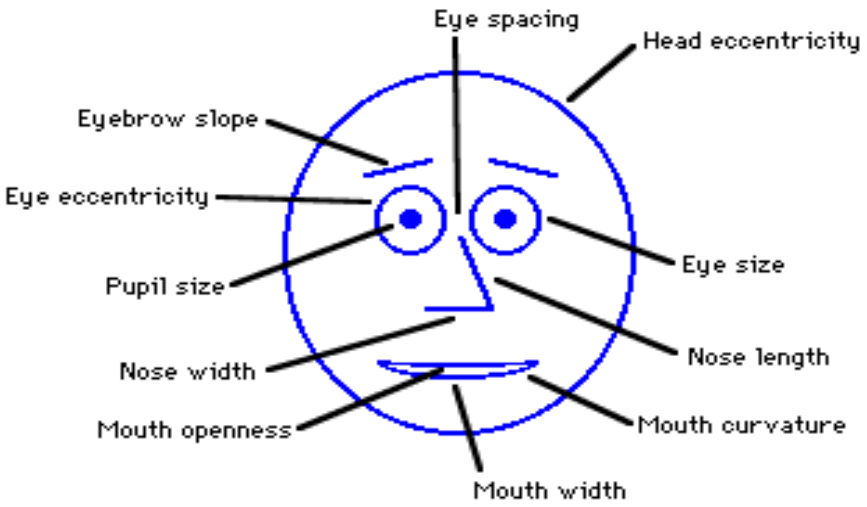


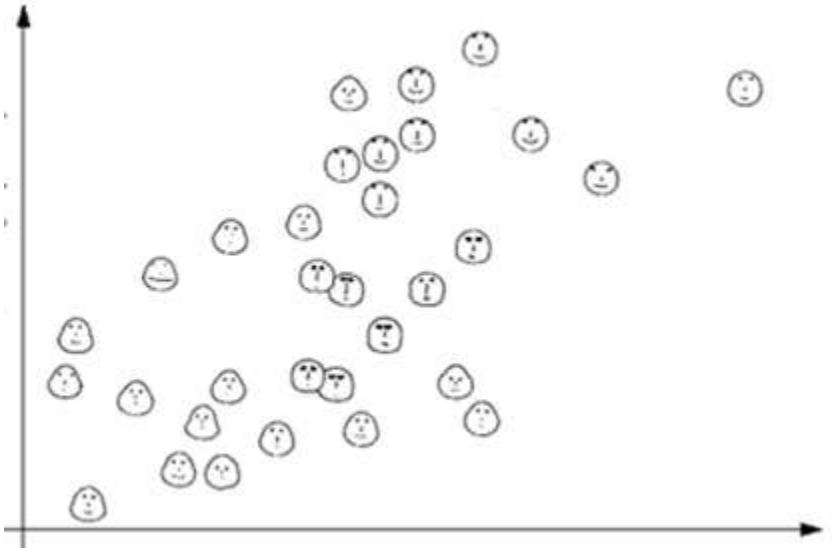
Паралелна визуализация

- Всяка стойност (точка) се асоциира с линия
- Подобни точки лежат на подобни линии
- Интерактивно изследване и сегментиране
- Недостатък
  - до 20 атрибути

Chernoff Faces

Кодиране на стойностите на различни променливи чрез характеристики на човешко лице

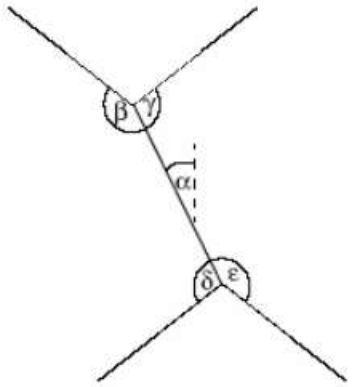




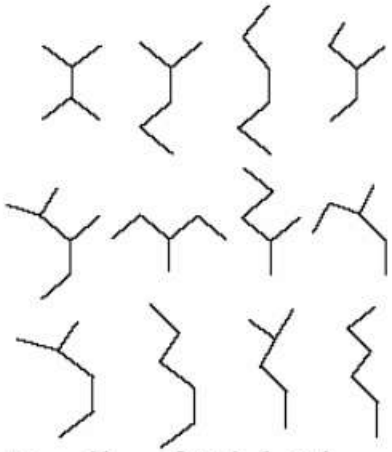
Cute applets: http://www.cs.uchicago.edu/~wiseman/chernoff/
http://hesketh.com/schampeo/projects/Faces/chernoff.html

Stick Figures

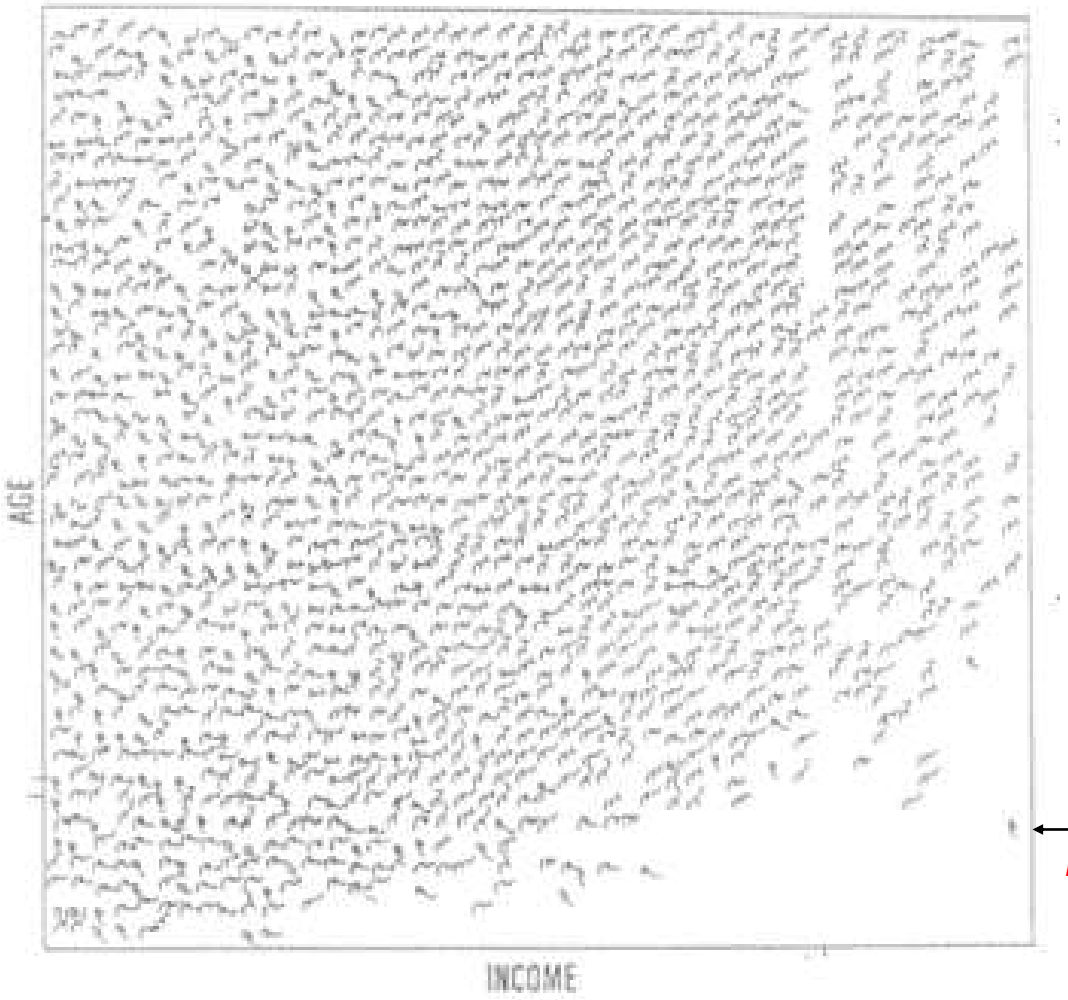
- Две променливи са избрани за координати X,Y
- Другите съответстват на различни дължини и ъгли
- Получената картина информира



Stick Figure Icon


A Family of Stick Figures

Stick Figures, пример



Данни от преброяване на населението:

- възраст
- образование
- пол

Една млада жена с високи доходи