

# ESMA 4016

# Data Visualization

Dr. Edgar Acuna

Departamento de Ciencias Matematicas

Universidad de Puerto Rico Recinto de Mayaguez

Website: [academic.uprm.edu/eacuna](http://academic.uprm.edu/eacuna)

# Contenido

- Uso de Visualizacion
- Representando datos en 1,2, y 3-D
- Representando datos en mas de 4 dimensiones
  - Scatterplot Matrix
  - Survey plots
  - Parallel coordinates
  - Radviz, Starcoord

# El uso de Visualizacion

- Visualizacion es el proceso de transformar la informacion en una forma visual de tal manera que el usuario pueda observar toda la informacion.
- El uso de una buena tecnica de visualizacion en data mining puede reducir el tiempo que toma entender los datos, encontrar relaciones entre las variables y descubrir informacion.
- Uno de los objetivos de la visualizacion es hacer analisis exploratorio de los datos.

# El uso de visualization (cont)

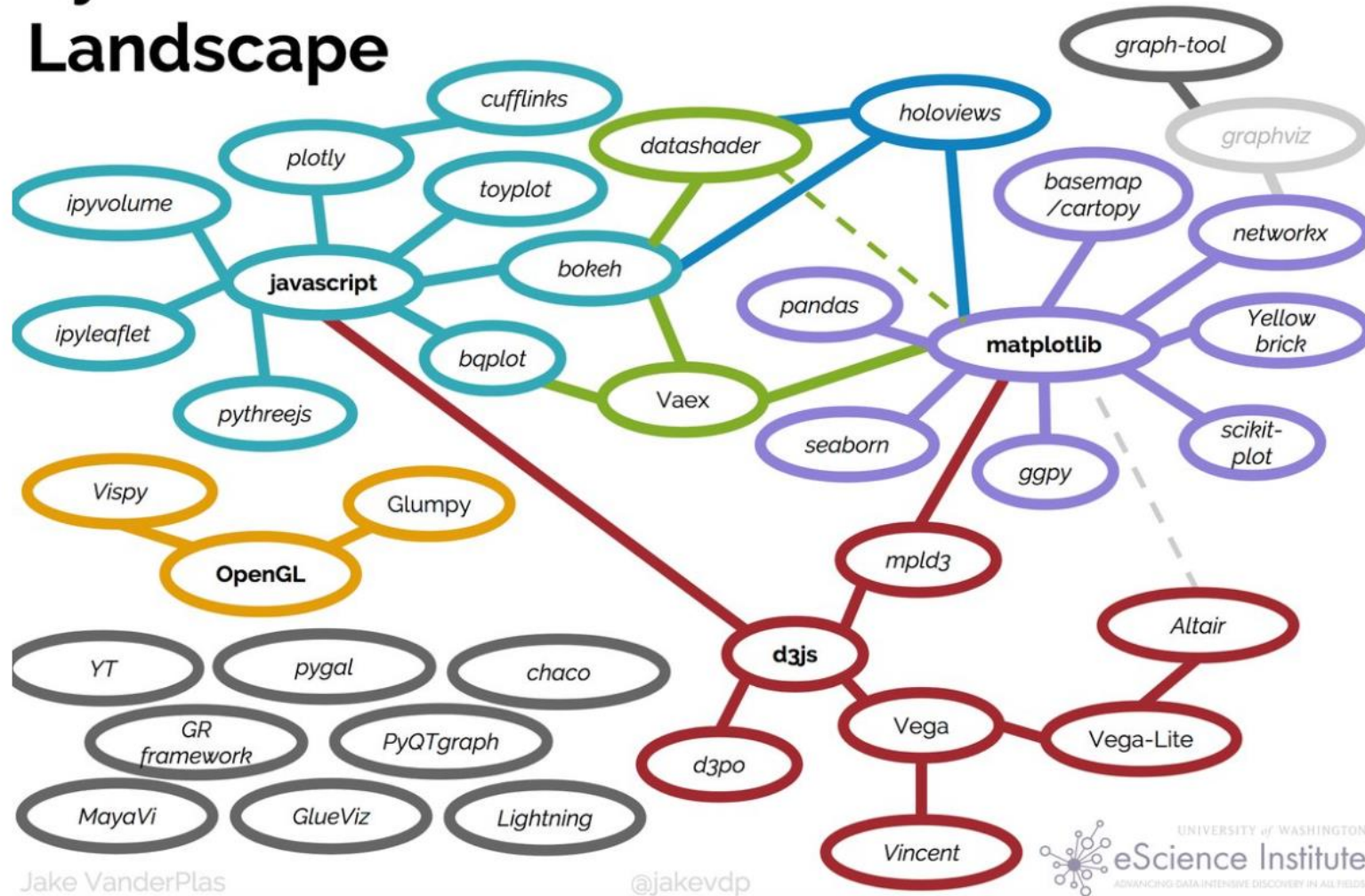
- En analisis exporatorio, se usan tecnicas de visualizacion antes de aplicar un algoritmo de data mining para obtener informacion de las características del dataset. El resultado de la exploracion puede conducir a formular hipotesis acerca de los datos.
- El uso de visualizacion permite al usuario mejorar el entendimiento de sus datos y evitar que pueda cometer errores en sus conclusions.
- Ayuda en la presentacion de los resultados
- Desventaja: Requiere de los ojos del humano y puede ser mal interpretada

# Los principios de Tufte de una Buena grafica

- Dar al observador
  - El mayor numero de ideas
  - En el tiempo mas corto
  - Con el minimo de Tinta en el espacio mas pequeno.
- Decir la verdad acerca de los datos!

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

# Python's Visualization Landscape



# Metodos de visualizacion

- Visualizando en 1-D, 2-D y 3-D
  - Hay bastantes metodos conocidos
- Visualizando en mas dimensiones
  - Scatterplot matricial
  - Survey plots
  - Parallel Coordinates
  - Radviz
  - Star Coordinates

# 1-D (Univariate) data

Python:

Dotplot esta en la libreria plotnine

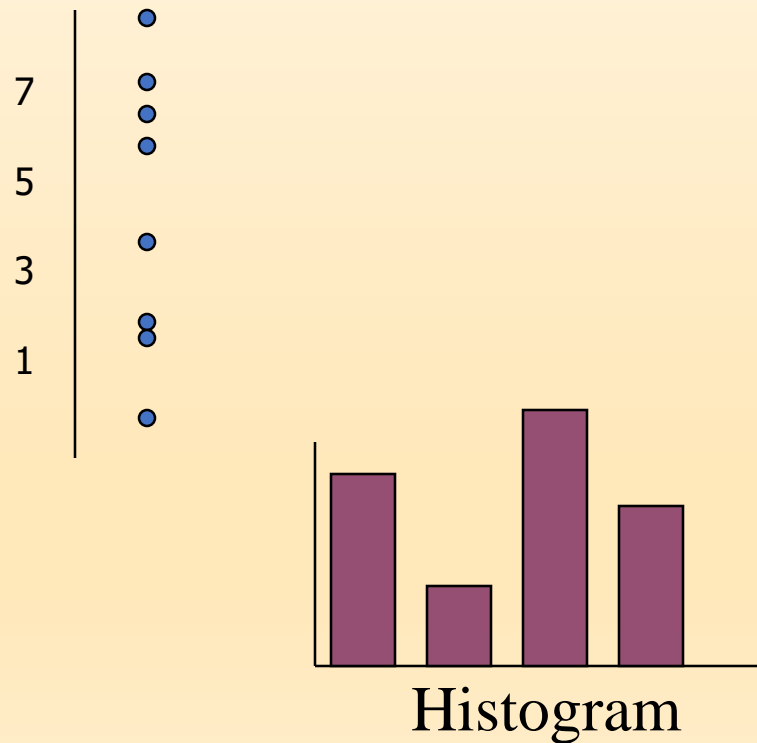
Histograma en varias librerias: plotnine, matplotlib, seaborn, plotly, bokeh

Boxplot en plotnine, matplotlib, seaborn, plotly y bokeh

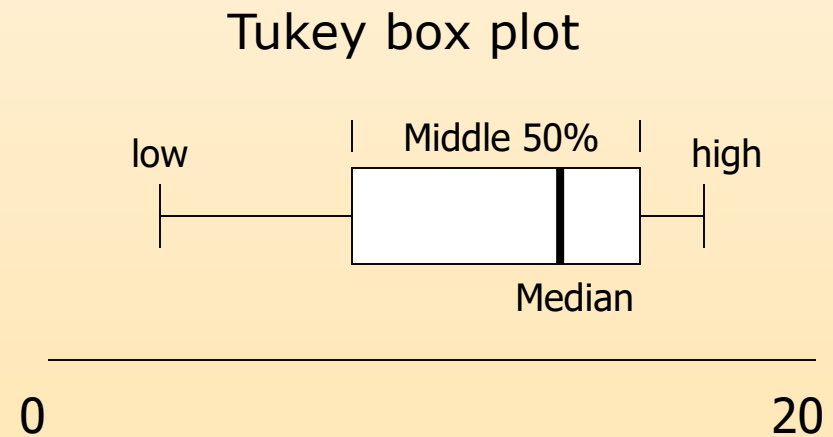


# 1-D (Univariate) Data

- Representations



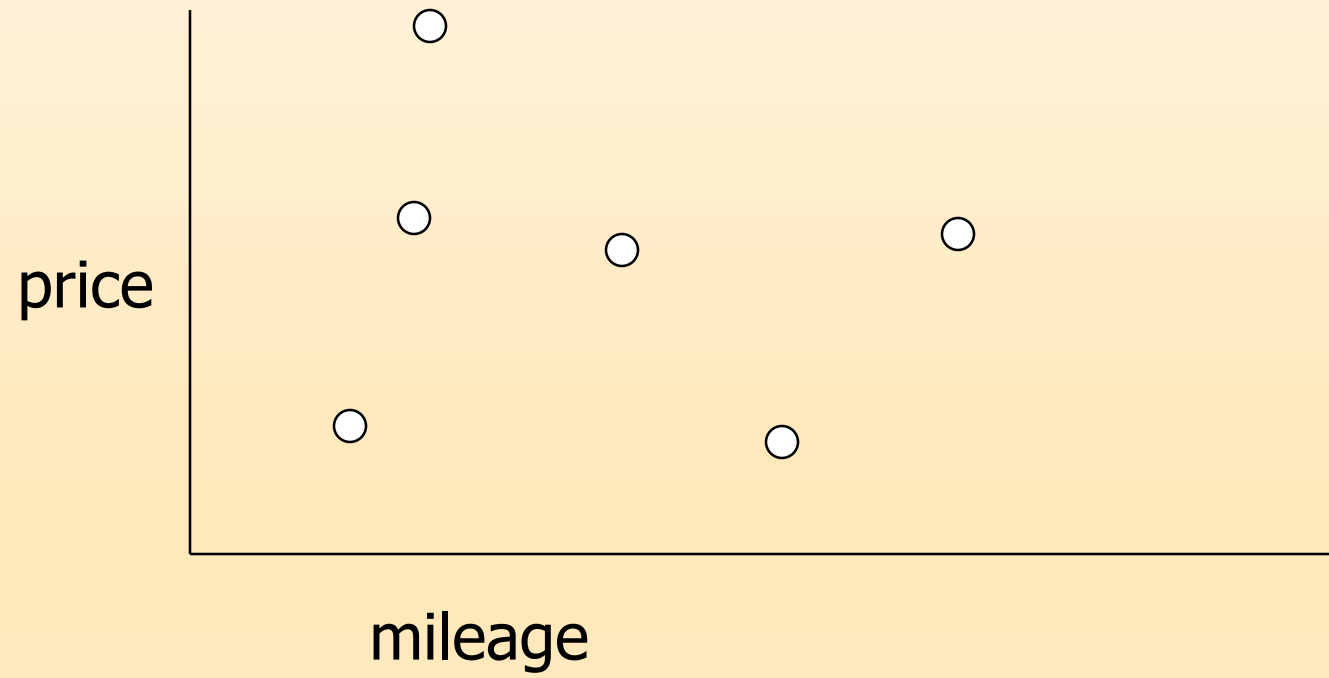
ESMA 4016



Edgar Acuna

# 2-D (Bivariate) Data

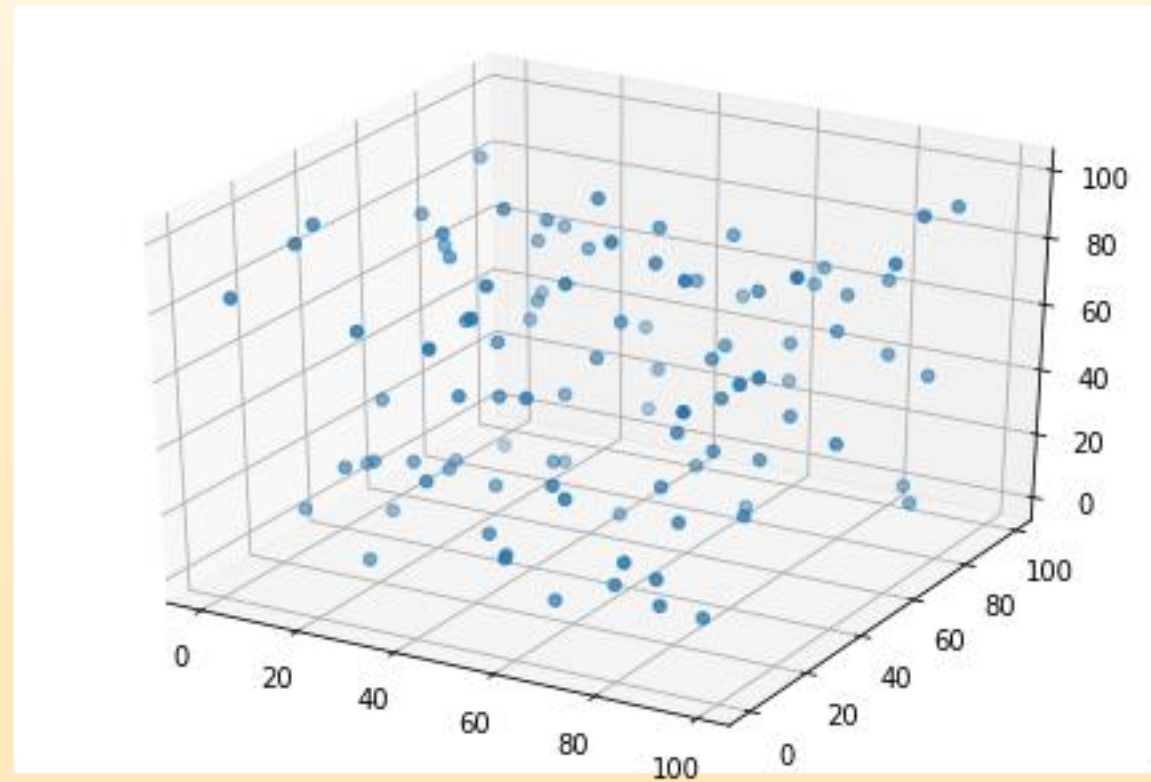
- Scatter plot, ...



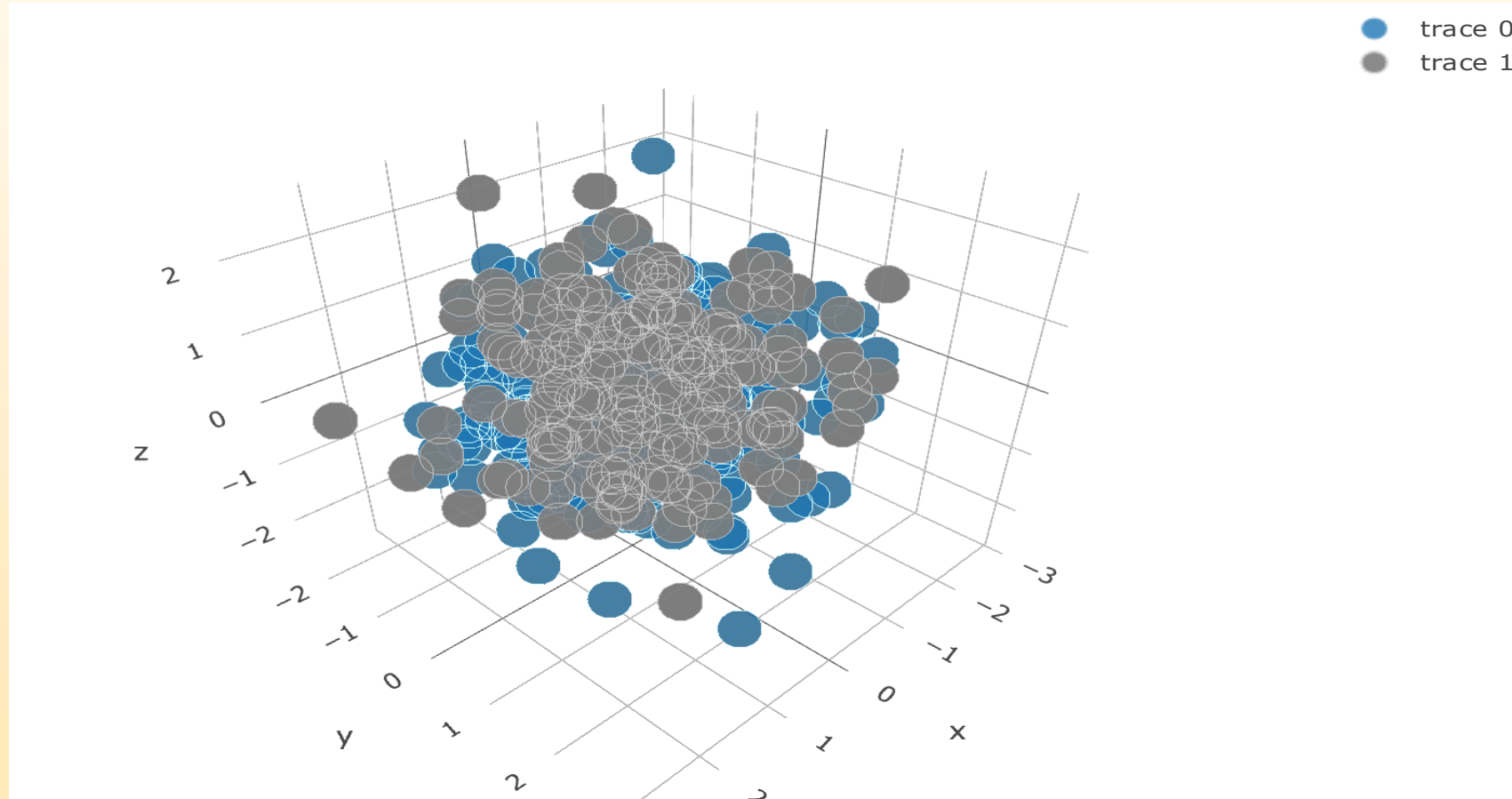
# 3-D Data

- En Python: Las libreria matplotlib, bokeh, plotly hacen scatterplot 3D

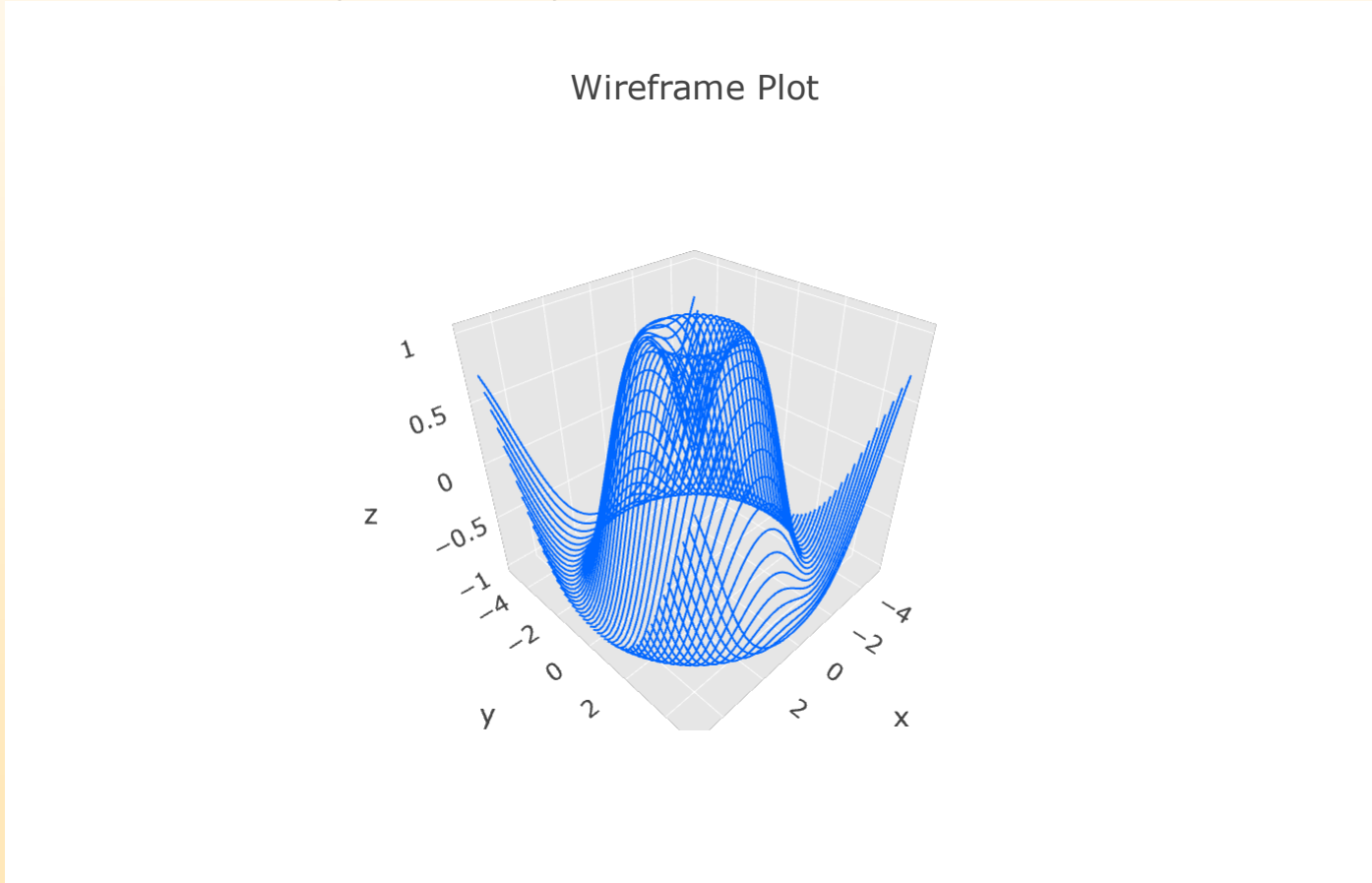
# Scatter3d using matplotlib



# Scatterplot 3d using plotly



# 3-D wireframe(plotly)



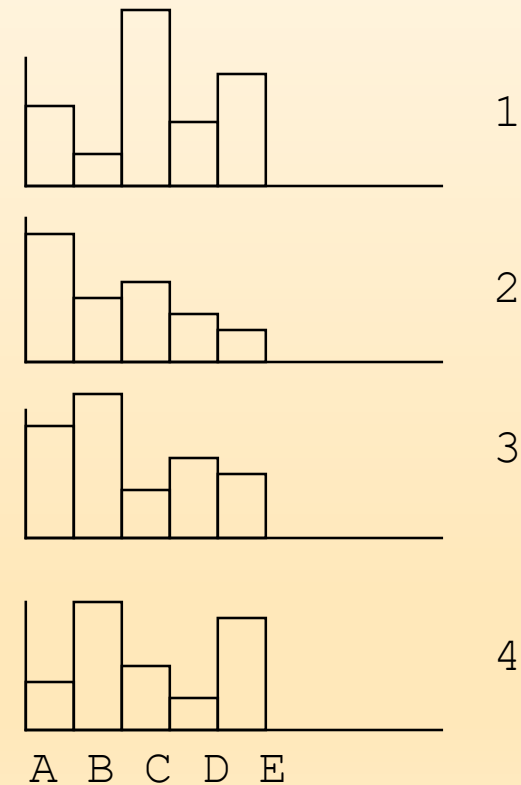
# Visualizando en 4 o mas Dimensiones

- Scatterplot Matrix
- Stars plot
- Parallel coordinate plot
- Radviz
- Star Coordinates

# Multiple Vistas

Cada variable es graficada separadamente

	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



Problema: No se muestran las correlaciones



# Scatterplot Matrix

Presenta plots 2-D para cada possible par de variables.

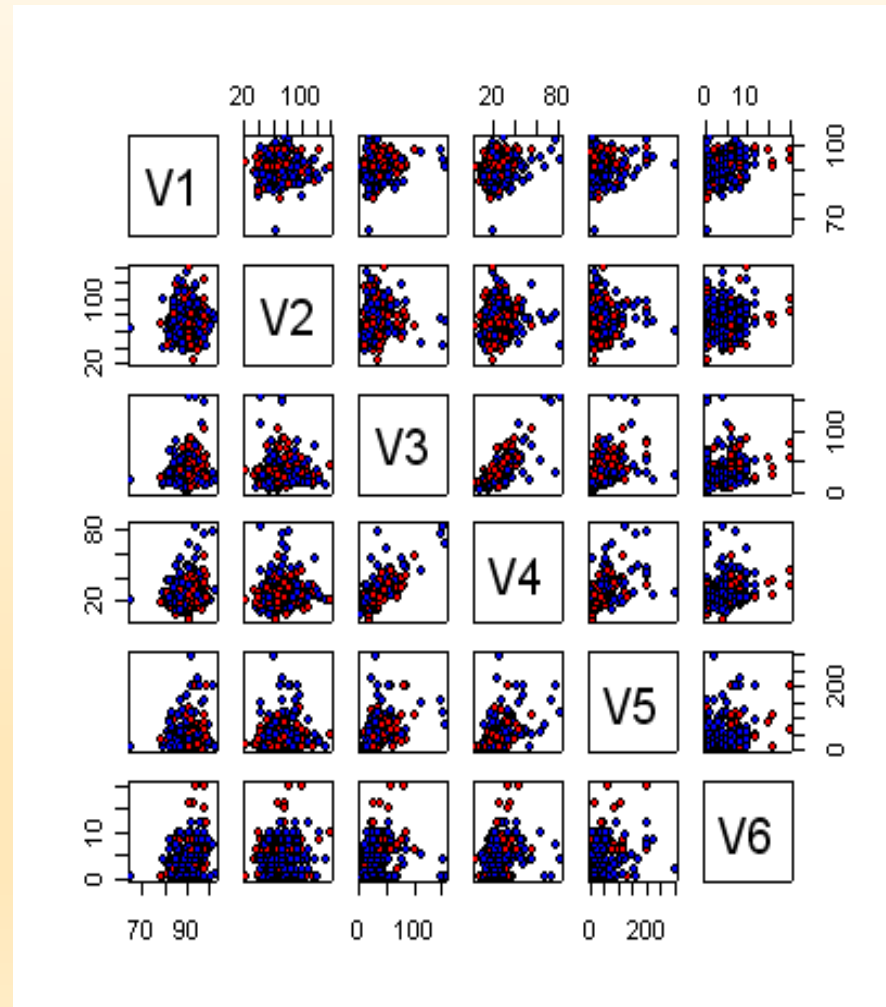
*Util para detectar*

Correlaciones Lineales  
(e.g. V3 & V4)

*Pero no detecta*

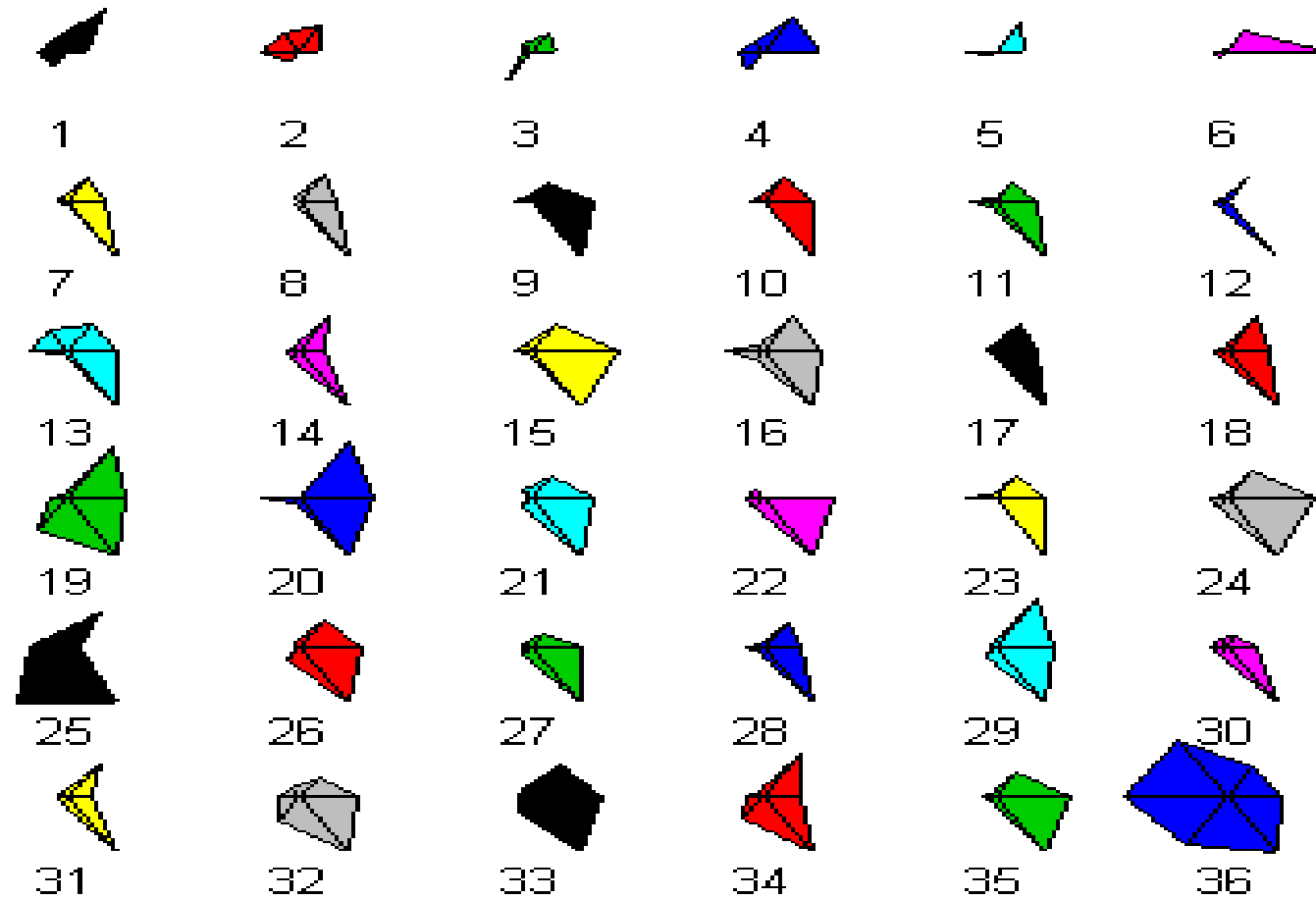
efectos multivariados

Esta disponible en Pandas,  
seaborn y en plotly.



# Stars Plots (Chambers et al., 1983)

**stars plot for bupa(instances 1:36)**



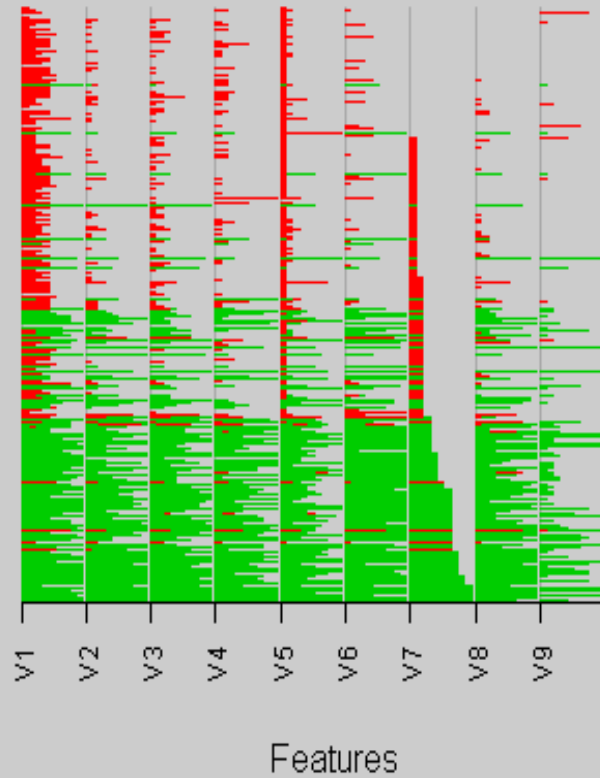
# The survey plot (Lohninger, 1994)

- A visualization invented by a French cartographer, Jacques Bertin, that is closely related to the visualization techniques: bar graph and permutation matrix.
- Consists of  $n$  rectangular areas or lines, one for each dimension of the dataset, that are vertically arranged.
- Each data value of an attribute is mapped to a point on the vertical line and the point is extended to a line with length proportional to the corresponding value.
- The strength of this visualization lays in its ability to show the relations and dependencies between any two attributes, especially when the data is sorted on a particular dimension.

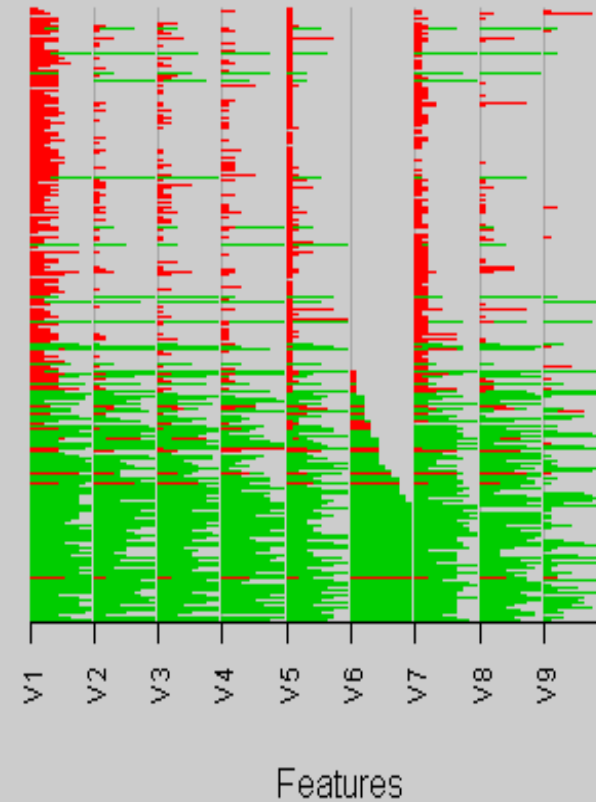
# The survey plot:

surveyplot(dataset: matrix , name: string, class: integer,  
orderon: integer, obs: list of integer )

**Survey Plot for breastw(ordered by attr=7**

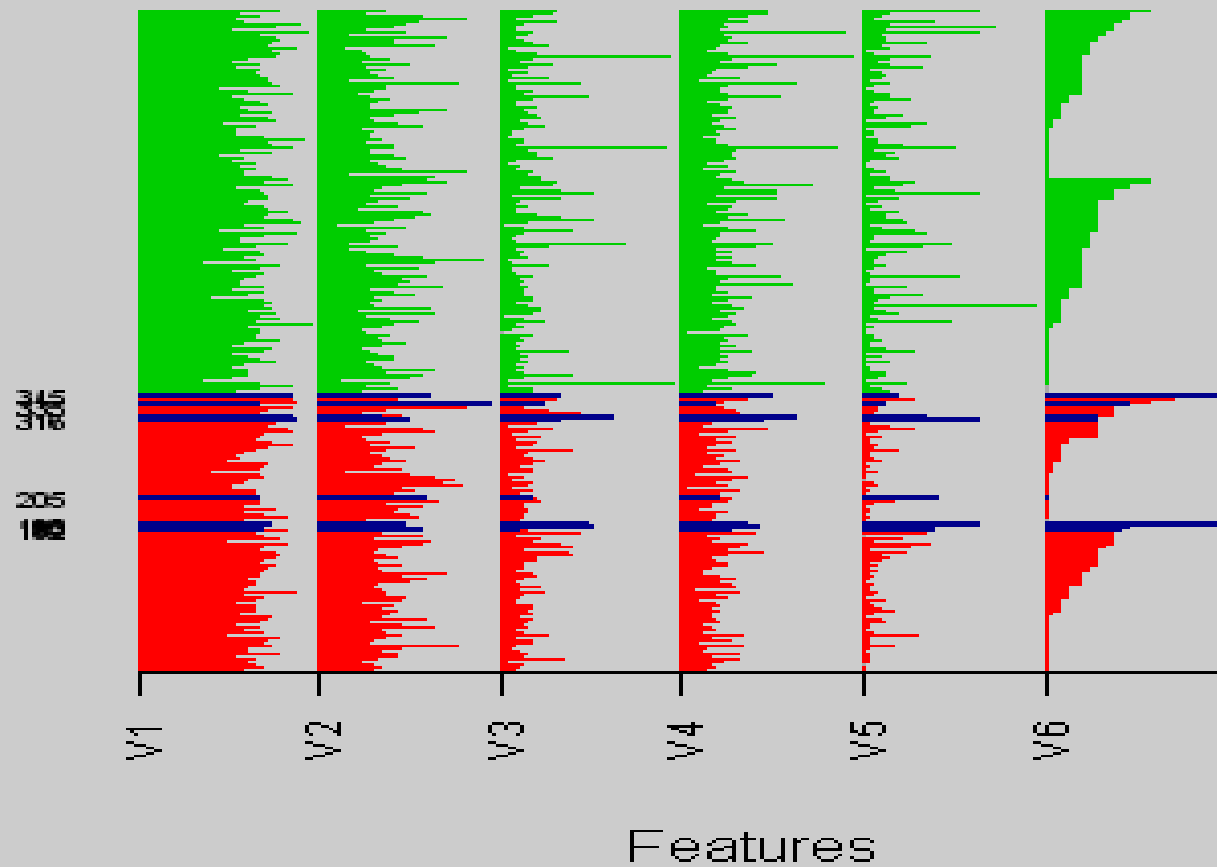


**Survey Plot for breastw(ordered by attr=6**



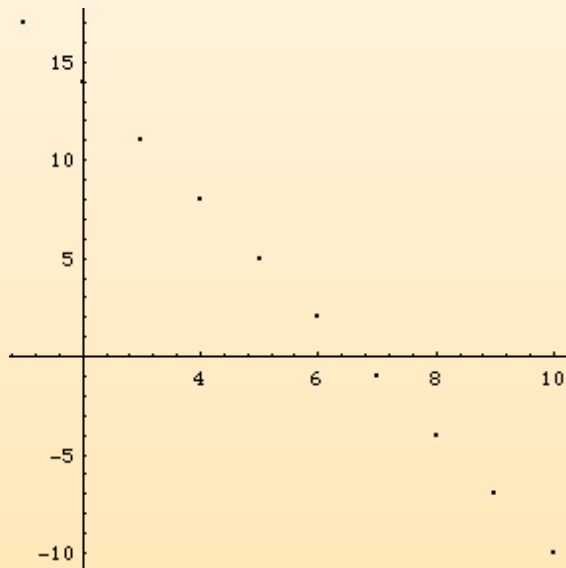
# Surveyplot as a tool to detect outliers

**Survey Plot for bupa(outliers class 1)**

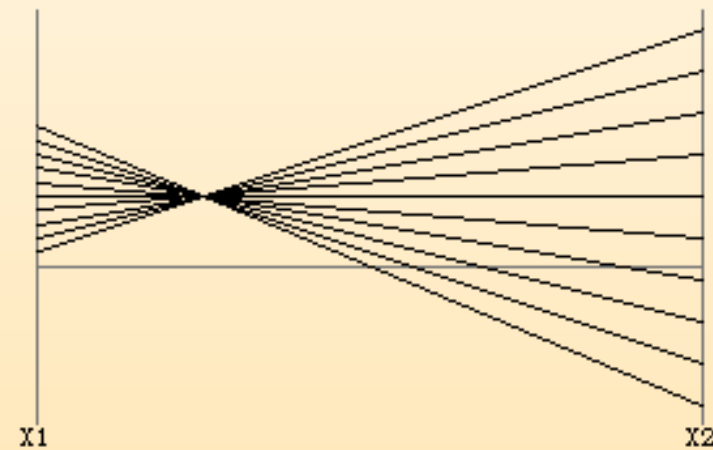


# Parallel Coordinates

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



Dataset in a Cartesian coordinates



Same dataset in parallel coordinates

Invented by Alfred Inselberg  
while at IBM, 1985

# The parallel coordinate plot:

- The parallel coordinate plot, described by Al Inselberg (1985), represents multidimensional data using lines.
- Whereas in traditional Cartesian coordinates all axes are mutually perpendicular, in parallel coordinate plots, all axes are parallel to one another and equally spaced.
- In this approach, a point in  $m$ -dimensional space is represented as a series of  $m-1$  line segments in 2-dimensional space. Thus, if the original data observation is written as  $(x_1, x_2, \dots, x_m)$ , then its parallel coordinate representation is the  $m-1$  line segments connecting points  $(1, x_1)$ ,  $(2, x_2)$ ,  $\dots$ ,  $(m, x_m)$ .
- Typically, features will be standardized before a parallel coordinate plot is drawn.

# Example: Visualizing Iris Data



Iris setosa

sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
...	...	...	...
5.9	3	5.1	1.8



Iris versicolor



Iris virginica



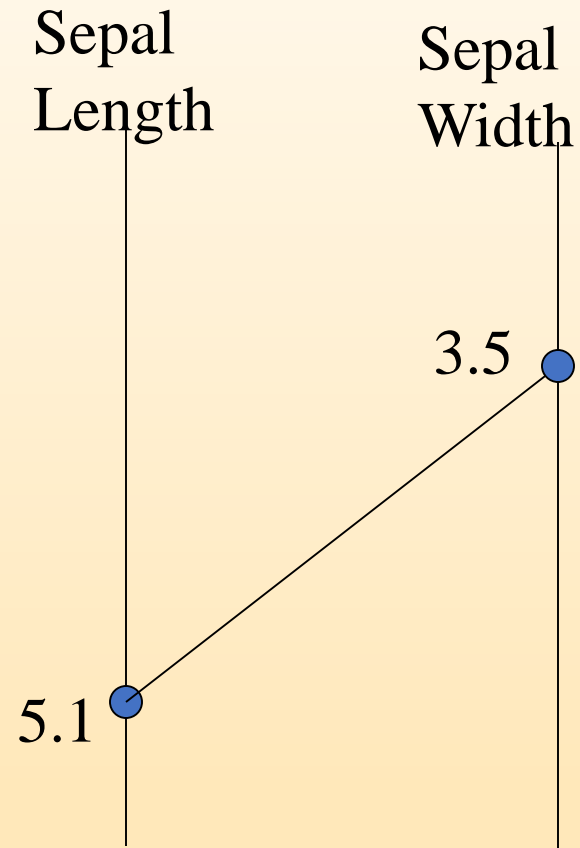
# Parallel Coordinates

Sepal  
Length

5.1

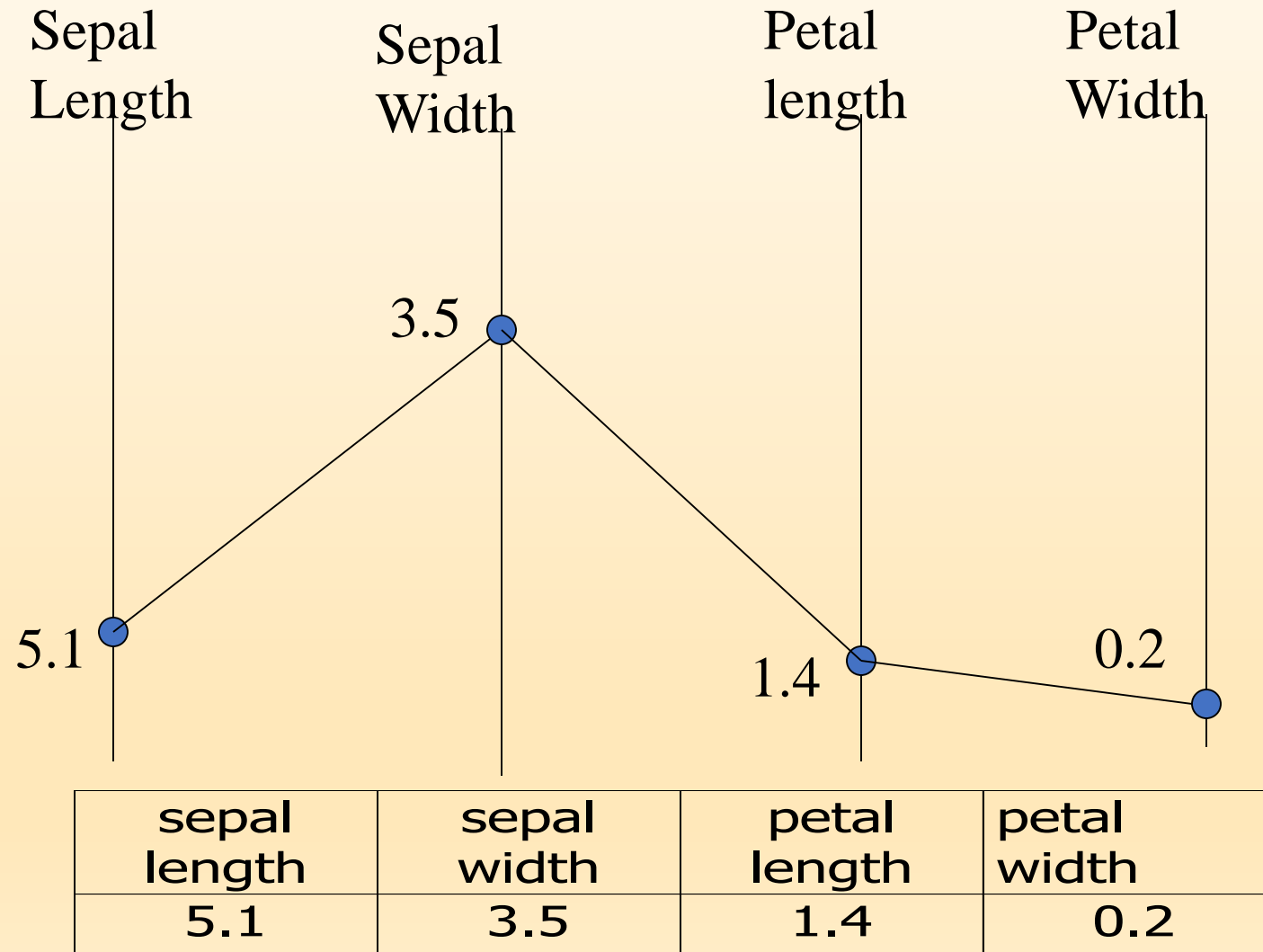
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

# Parallel Coordinates: 2 D

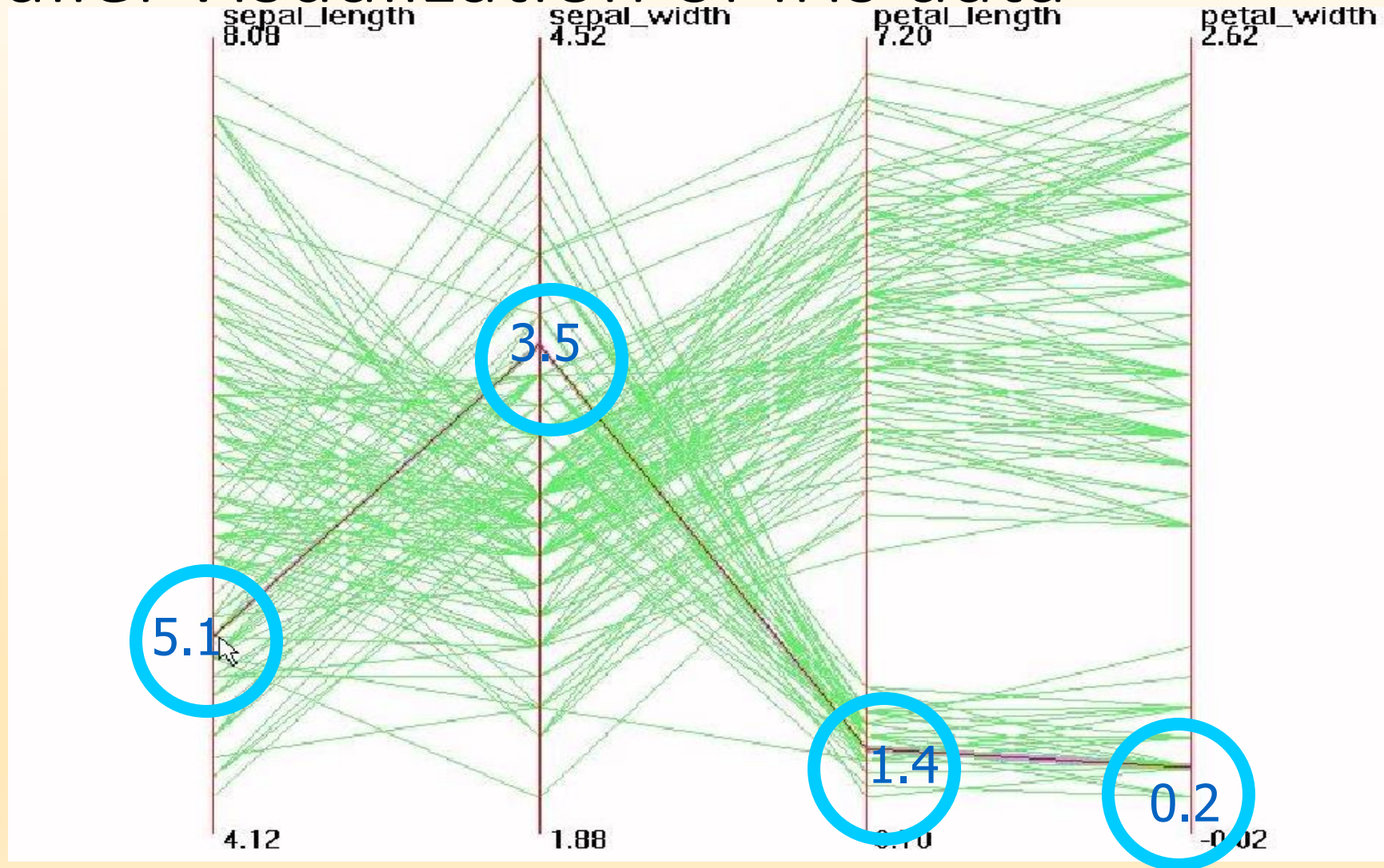


sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

# Parallel Coordinates: 4 D



# Parallel Visualization of Iris data



# Parallelplot (cont)

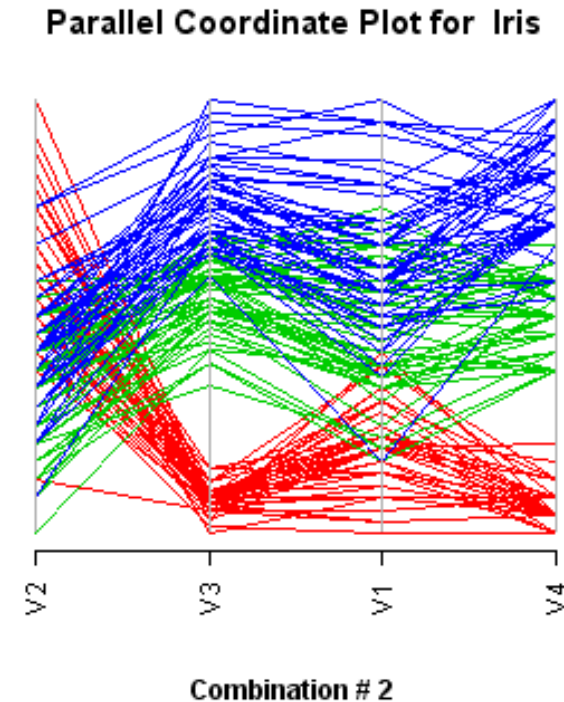
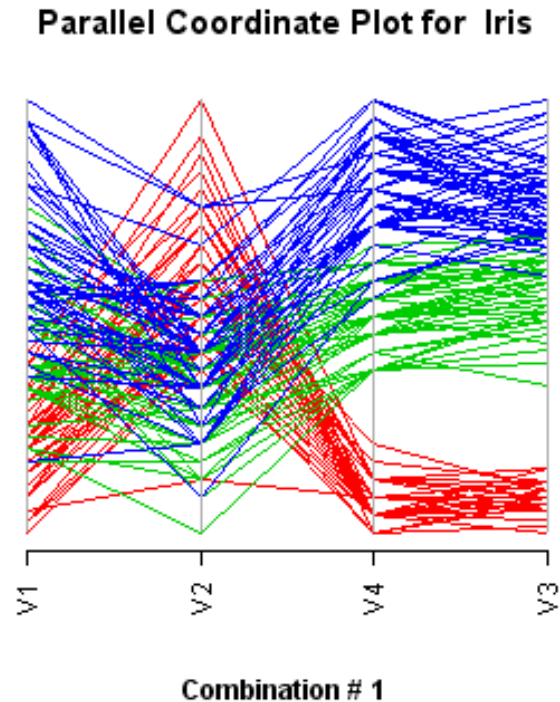
- Pairwise comparison is limited to those axis that are adjacent.
- For a dataset with  $p$  attributes there are  $p!$  permutations of the attributes so each of them is adjacent to every attribute in some permutation.
- Wegman (1990) determined that only  $\lfloor (p+1)/2 \rfloor$  permutations are needed. ( $\lfloor . \rfloor$  is the greatest integer function).

# The parallel coordinate plot

parallelplot(dataset: matrix , name: string, class: integer,  
comb: integer, obs: list of integer )

## Iris dataset:

- Data on the flowers.
- 4 attributes (sepal length, sepal width petal length, and petal width,)
- 150 instances
- 3 classes (Setosa, Versicolor, Virginica)
- No missing values.



## Interpretation:

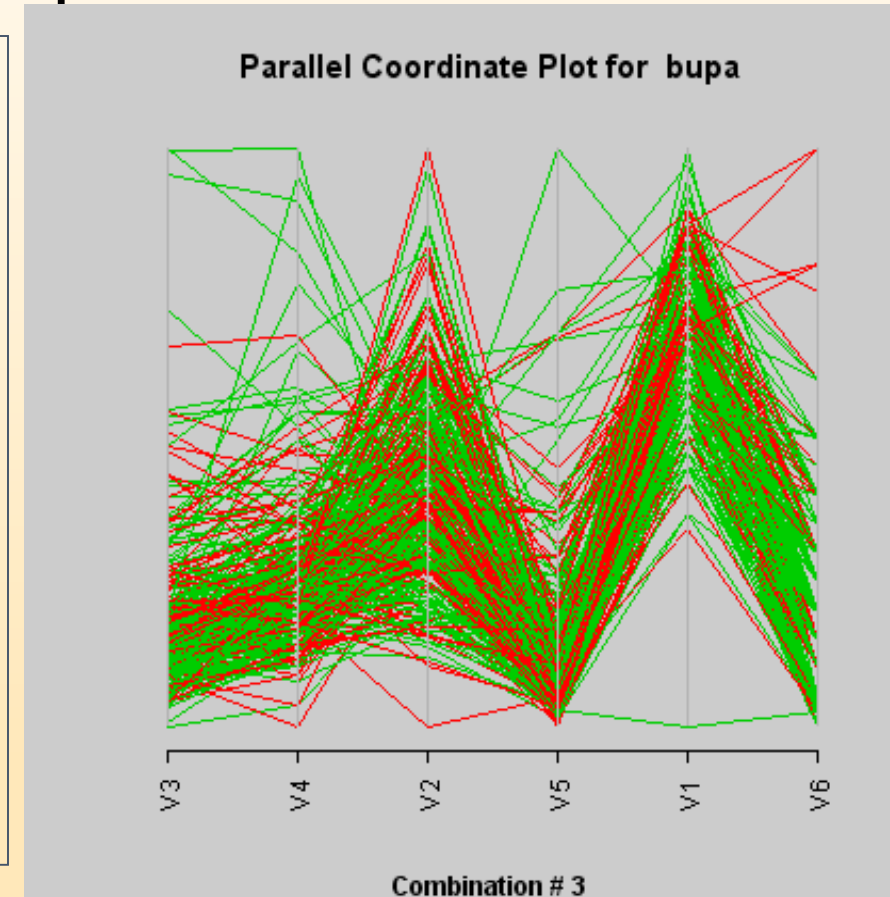
- Each different color represents a different class.
- If two attributes are highly positively correlated, lines passing from one feature to another tend not to intersect between the parallel coordinate axes.

# The parallel coordinate plot

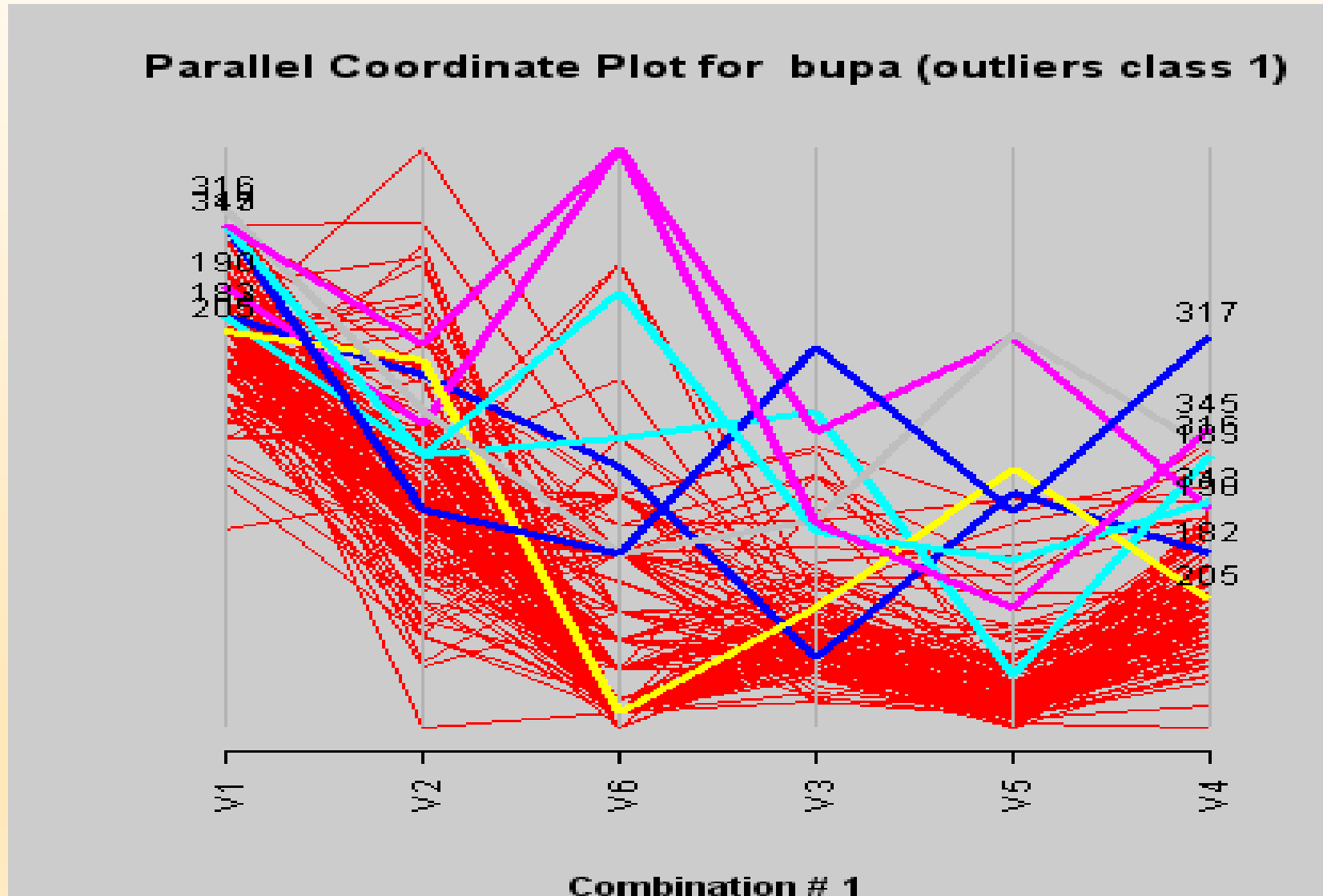
- **For highly negatively correlated attributes, the line segments tend to cross near a single point between the two parallel coordinate axes.**
- **The presences of outliers is suggested by poly-lines that do not follow the pattern for their class.**

**Some discrimination can be observed for several features.**

**One limitation of this displays is the loss of the information that is encoded into the lines between the axes for discrete, heterogeneous data attributes.**



# Parallelplot as a tool to detect outliers



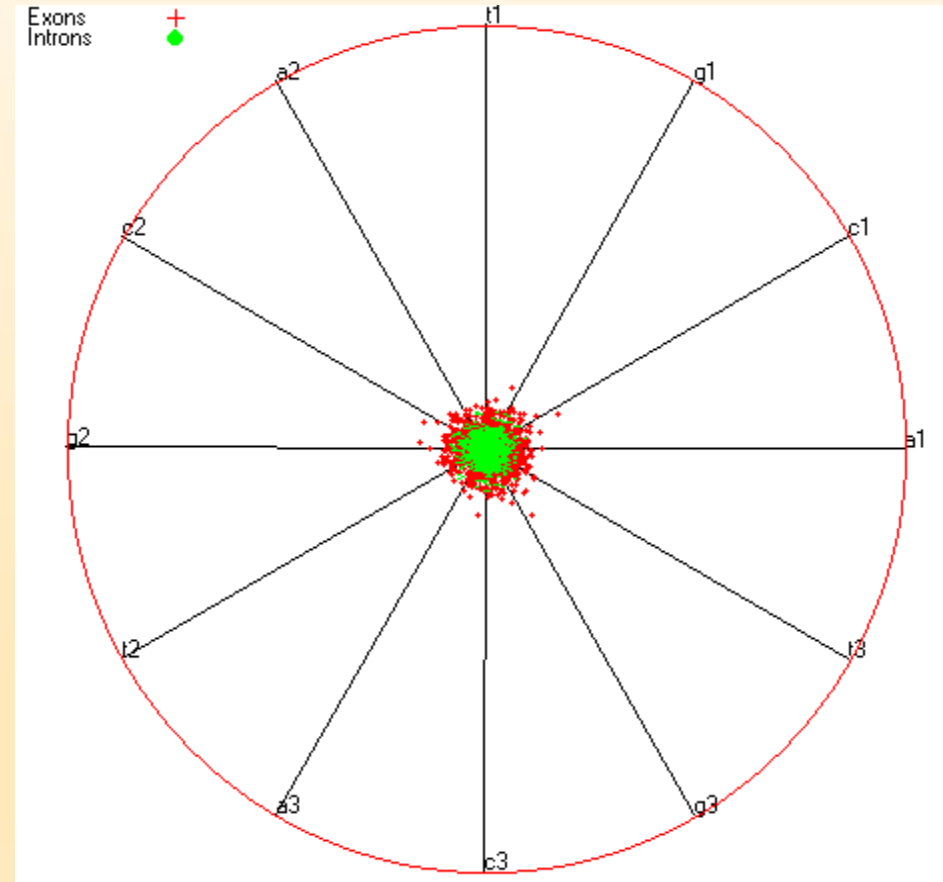


# Parallel Visualization Summary

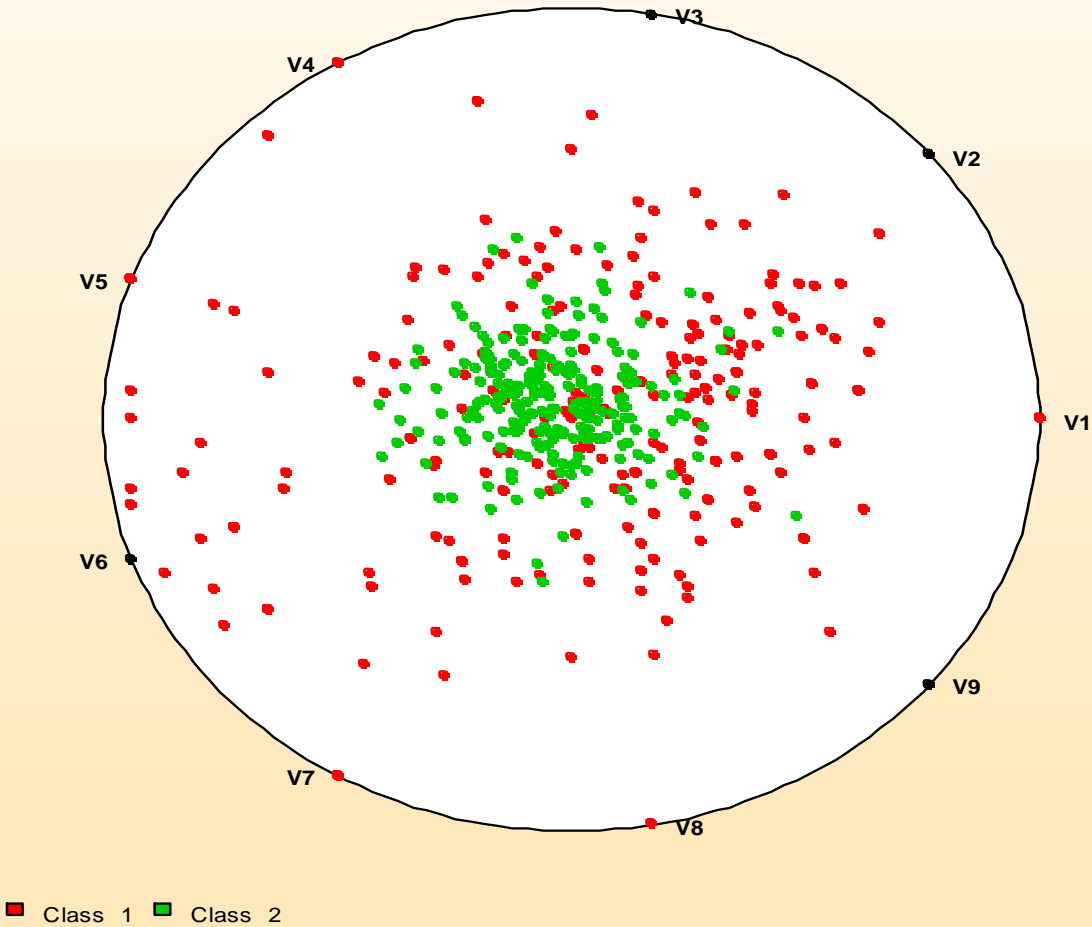
- Each data point is a line
  - Similar points correspond to similar lines
  - Lines crossing over correspond to negatively correlated attributes
  - Interactive exploration and clustering
- 
- Problems: order of axes, limit to ~20 dimensions
  - Esta disponible en Pandas

# RadViz (Ankerst, et al., 1996)

- a radial visualization
- One spring for each feature .
- One end attached to perimeter point where the feature position is located. The other end attached to a data point.
- Each data point is displayed inside the circle where the sum of the spring forces equals 0.
- Good for outlier detection
- Esta disponible en Pandas y en Orange

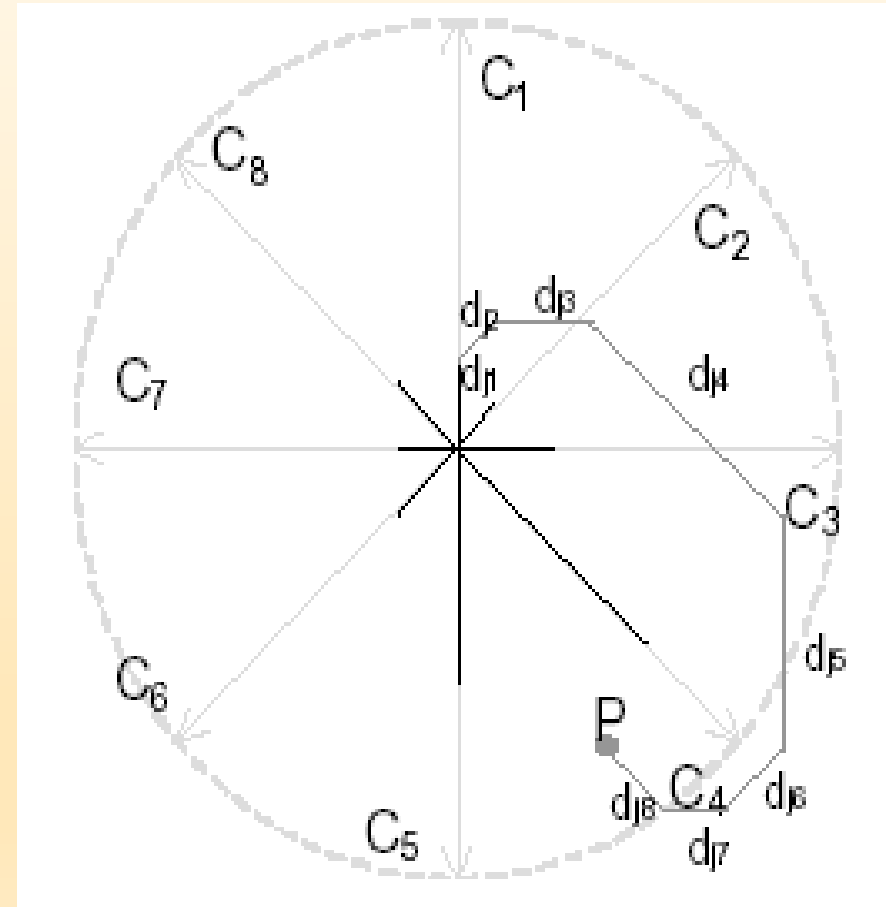


2D-Radviz for breastw



# Star Coordinates (Kandogan, 2001)

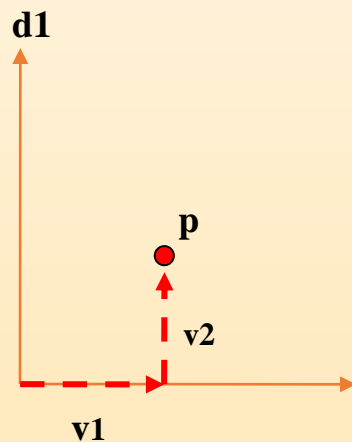
- Each dimension shown as an axis
- Data value in each dimension is represented as a vector.
- Data points are scaled to the length of the axis
  - min mapping to origin
  - max mapping to the end



# Star Coordinates Contd

## Cartesian

$$P=(v1, v2)$$

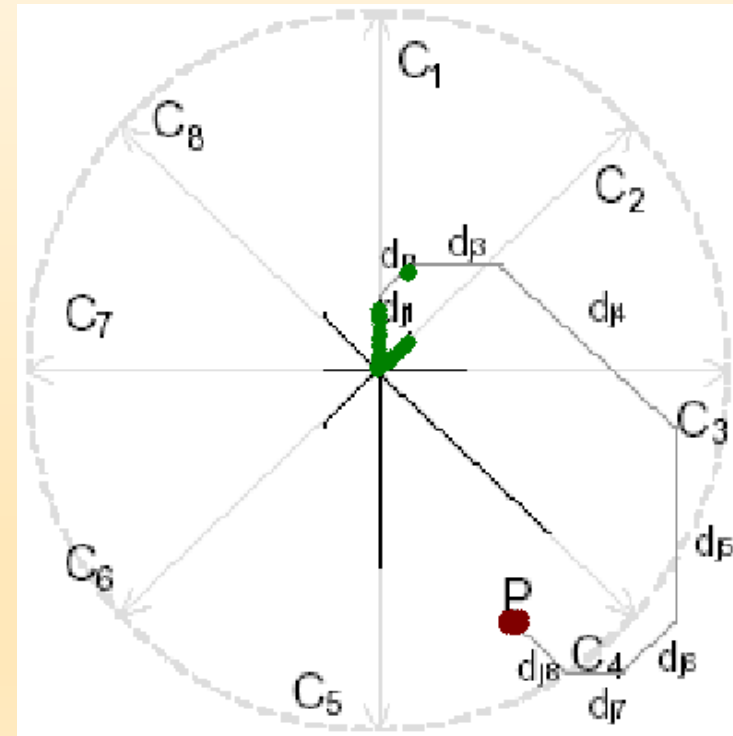


Mapping:

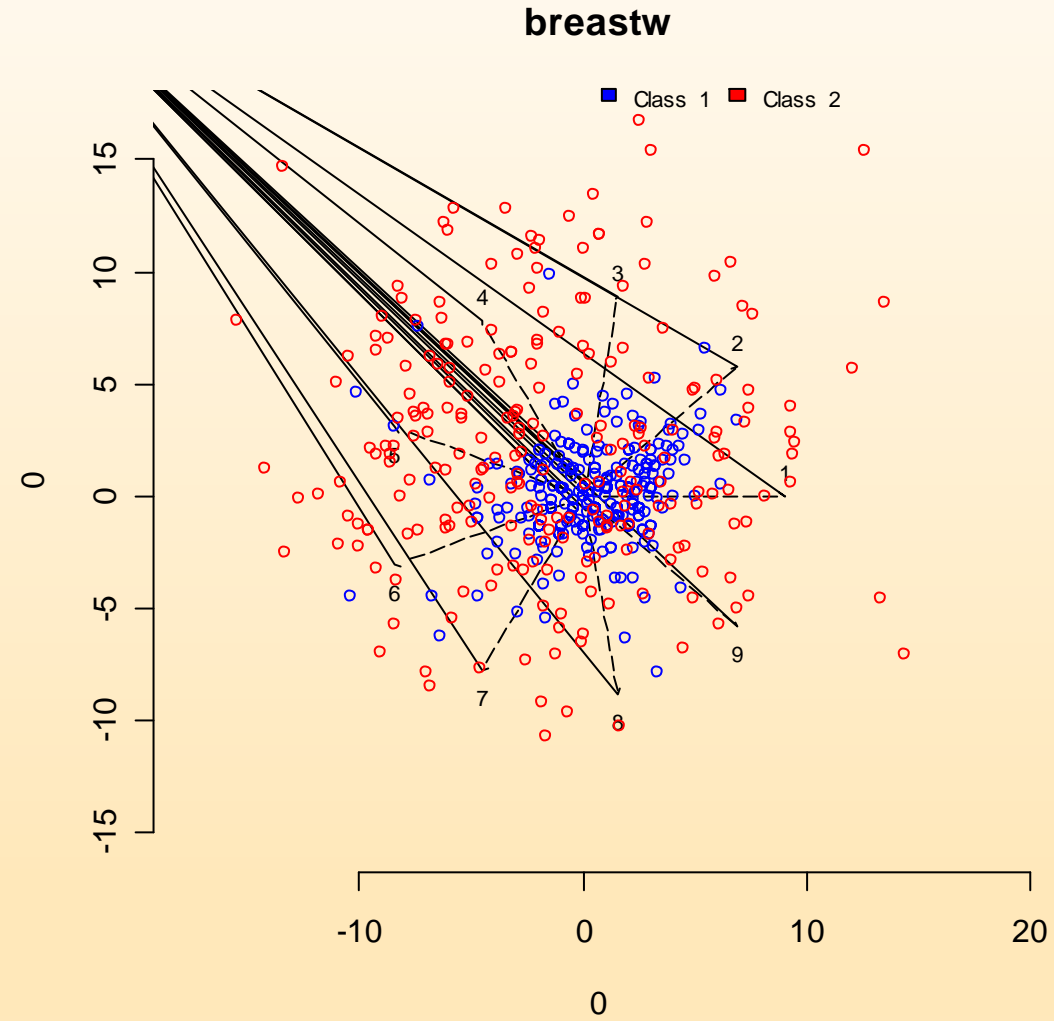
- Items  $\rightarrow$  dots
- $\Sigma$  attribute vectors  $\rightarrow$  position

## Star Coordinates

$$P=(v1,v2,v3,v4,v5,v6,v7,v8)$$



```
starcoord(breastw,main="breastw",class=T)
```



# Visualization software

## Free and Open-source

- Ggobi (before was xgobi). Built using Gtk. Interface with databases systems. Runs on Windows and Linux. <http://www.ggobi.org/>
- XmdvTool. The multivariate data visualization tool. Available for Linux and Windows. Built using OpenGL and Tcl/Tk. See <http://davis.wpi.edu/~xmdv/>
- Many more - see [www.kdnuggets.com/software/visualization.html](http://www.kdnuggets.com/software/visualization.html)