

# Visualización de datos

# Outline

- Buenos gráficos – el “lie factor”
- Representar datos en 1,2, y 3-D
- Representar datos en 4+ dimensiones
  - Parallel coordinates
  - Scatterplots
  - Stick figures

# Rol de la visualización

- Soporte de la exploración interactiva
- Ayuda a la presentación de resultados
- Contra: Puede llevar a confusiones, ser engañosas

# Una buena figura

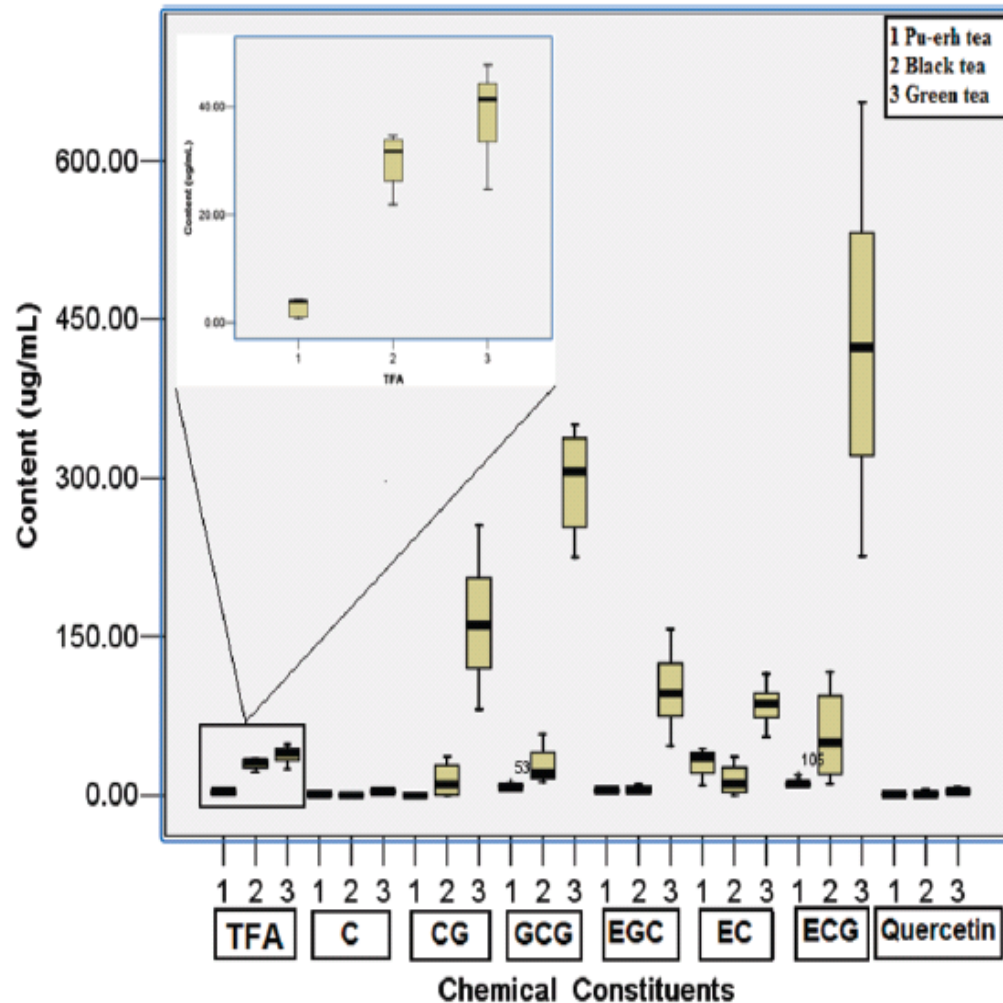
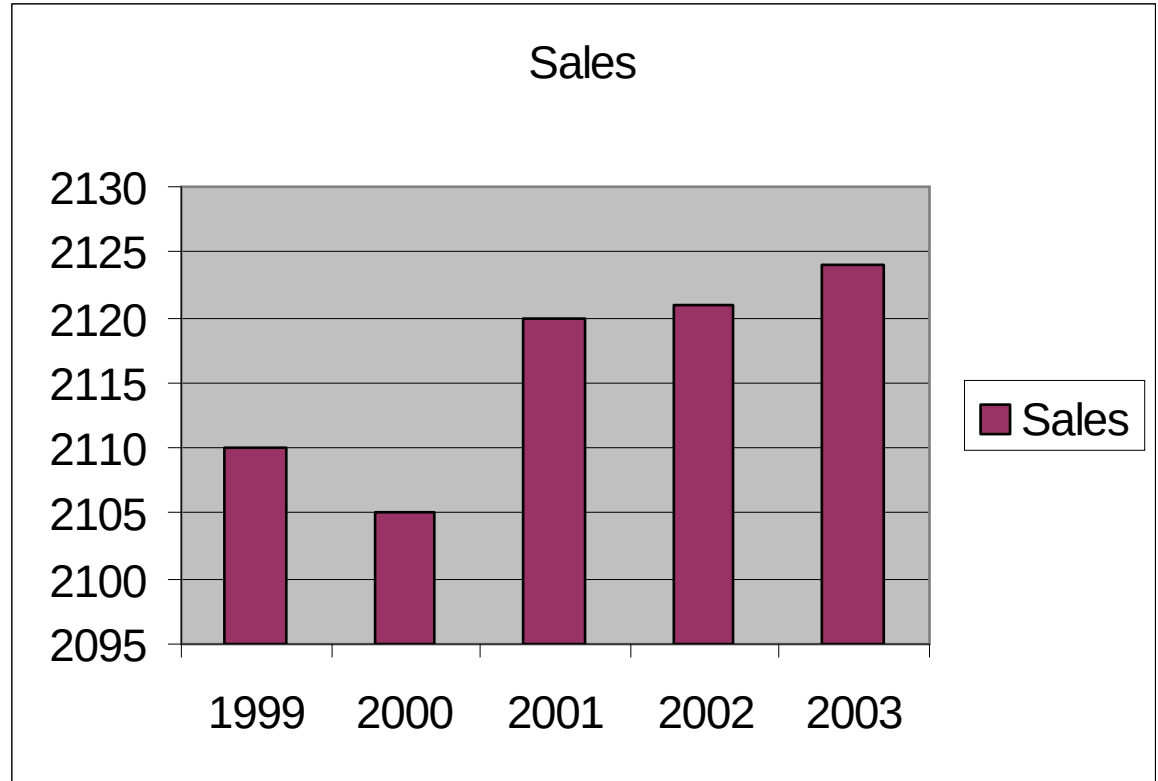


Figure 3. The contents of TFA, C, CG, GCG, EGC, EC, ECG, and quercetin in pu-erh teas, black teas, and green teas.

# Malas figuras: Spreadsheet

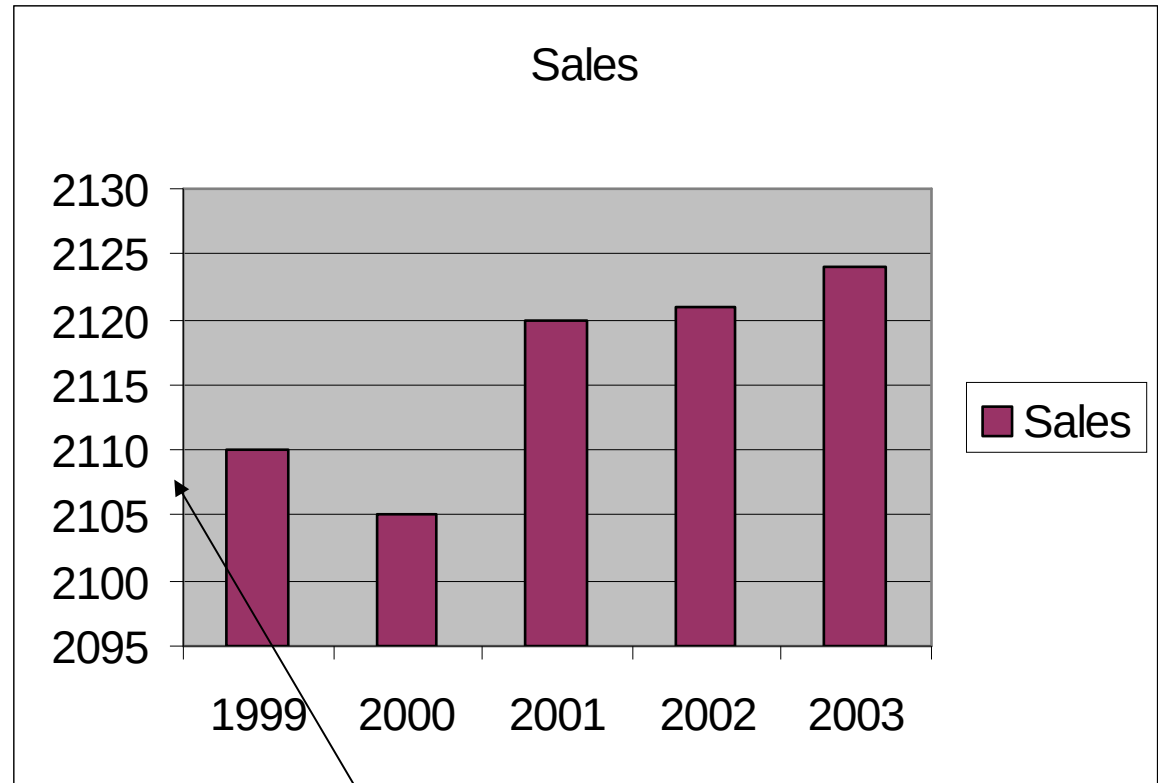
Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



Qué está mal en esta figura?

# Malas figuras: Spreadsheet con eje vertical engañoso

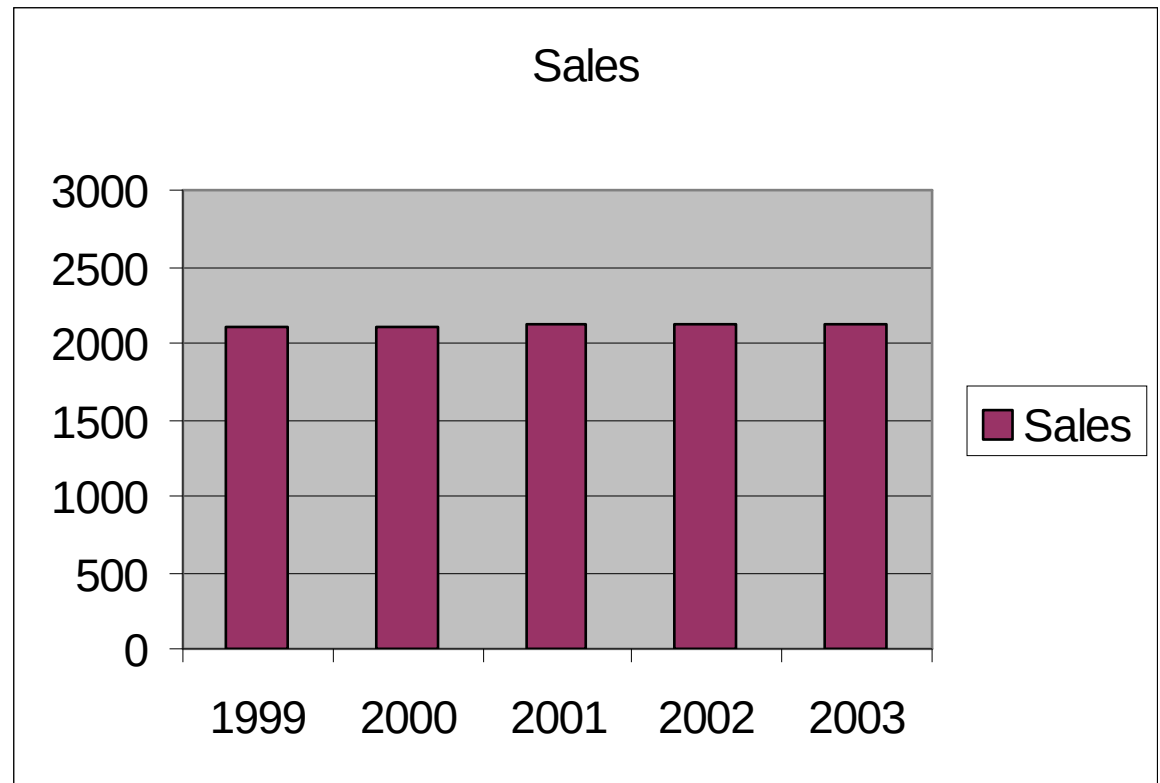
Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



La escala del eje Y da la impresión  
**INCORRECTA** de un gran efecto

# Una mejor figura

Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



La escala de 0 a 2000 da la correcta impresión de mínimo efecto

# Lie Factor

$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}} =$$

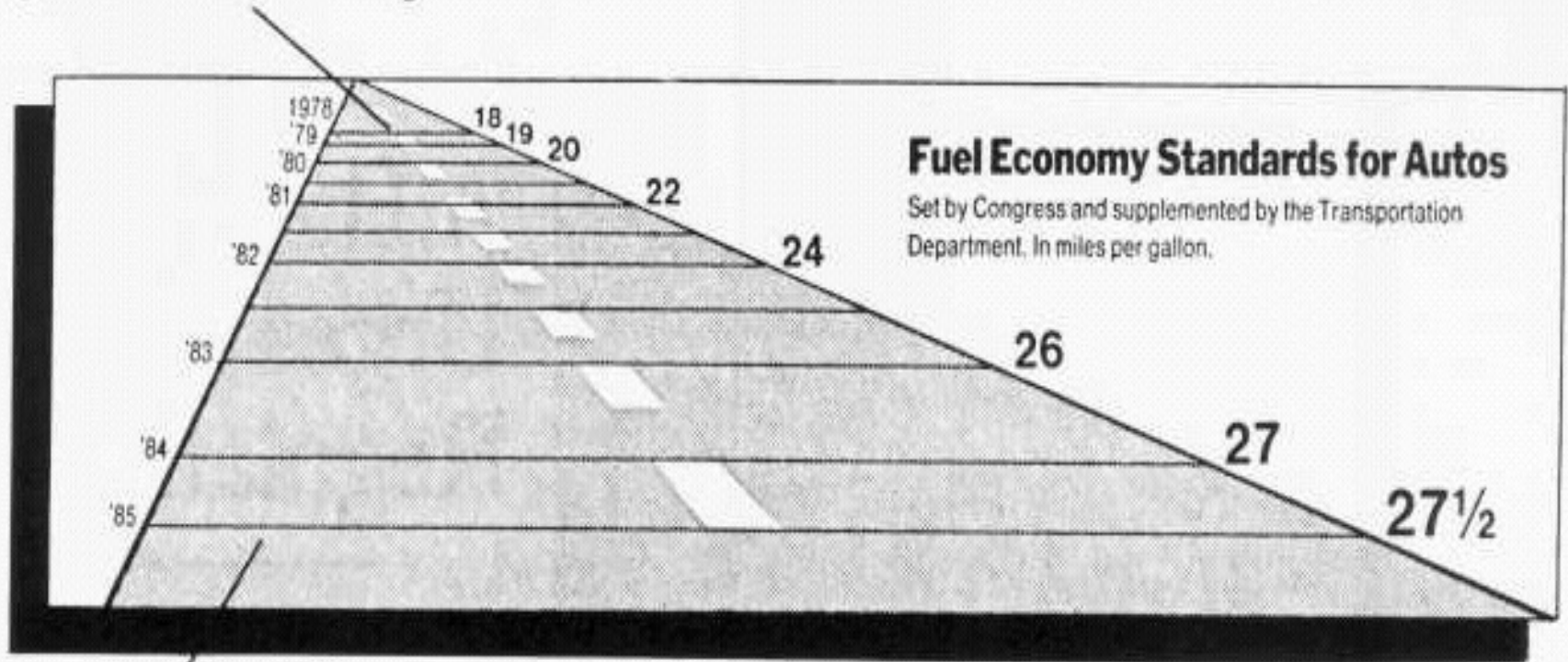
$$= \frac{\frac{(29-11)}{11}}{\frac{(2124-2105)}{2105}} = \frac{1.636}{0.009} = 181.2$$

Requerido por Tufte:  $0.95 < \text{Lie Factor} < 1.05$

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



This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

Lie Factor=14.8

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

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

# Tufte: Principios de las buenas visualizaciones

- Darle al observador
  - el mayor número de ideas
  - en el menor tiempo
  - con la menor cantidad de tinta y espacio.
- Decir la verdad sobre los datos!

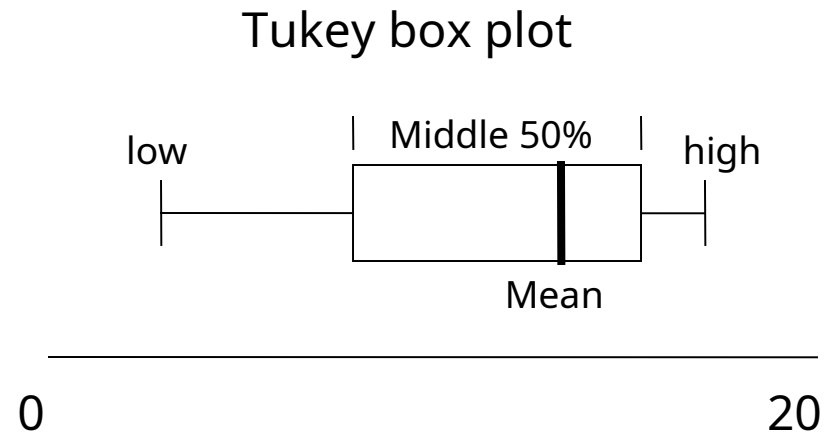
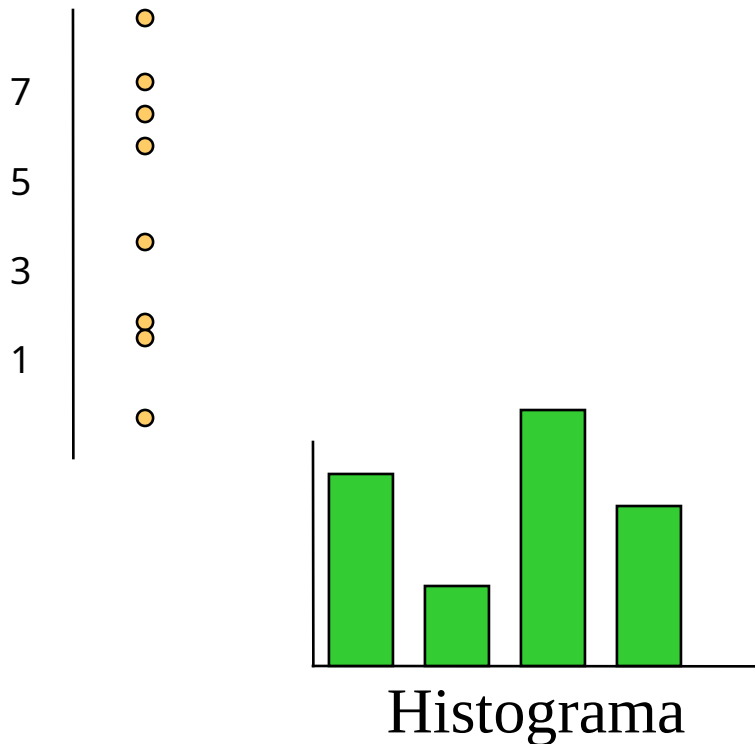
(E.R. Tufte, "The Visual Display of Quantitative Information", 2nd edition)

# Métodos de Visualización

- Visualización en 1-D, 2-D y 3-D
  - Métodos simples y conocidos
- Visualización en más dimensiones
  - Parallel Coordinates
  - Símbolos
  - Otras ideas

# Datos 1-D (Univariados)

## ■ Representaciones



# R: ejemplos

```
x<-c(10,runif(99))
```

```
plot(x)
```

```
plot(x,rep(0,100))
```

```
hist(x)
```

# se ve el outlier, pero en muchos datos se puede perder

```
x<-c(10,rnorm(9999))
```

```
hist(x)
```

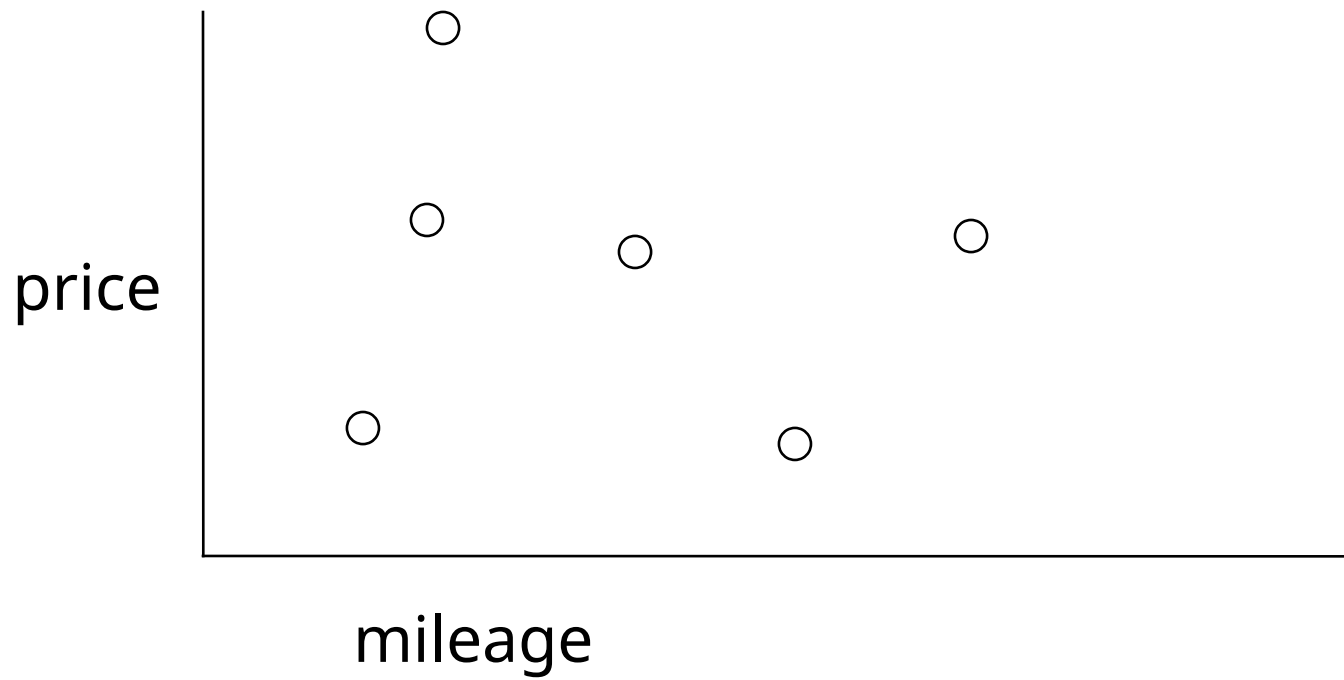
# con boxplot se ve el outlier siempre, pero se pierde información de la distribución

```
boxplot(x)
```

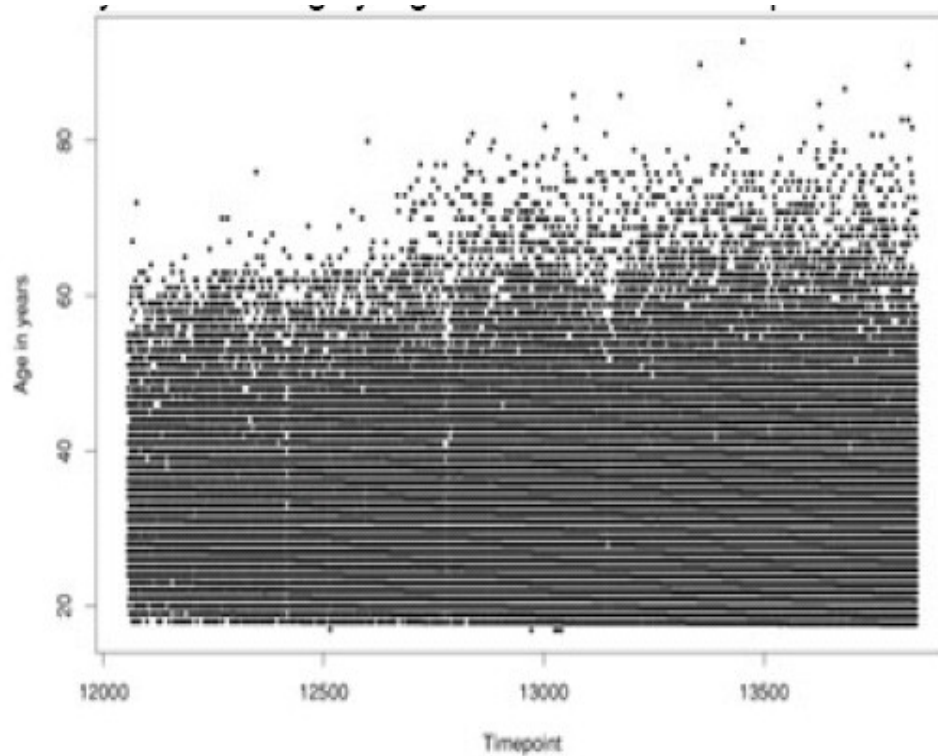
```
boxplot(x,range=2)
```

# Datos 2-D (Bivariados)

- Scatter plot, ...



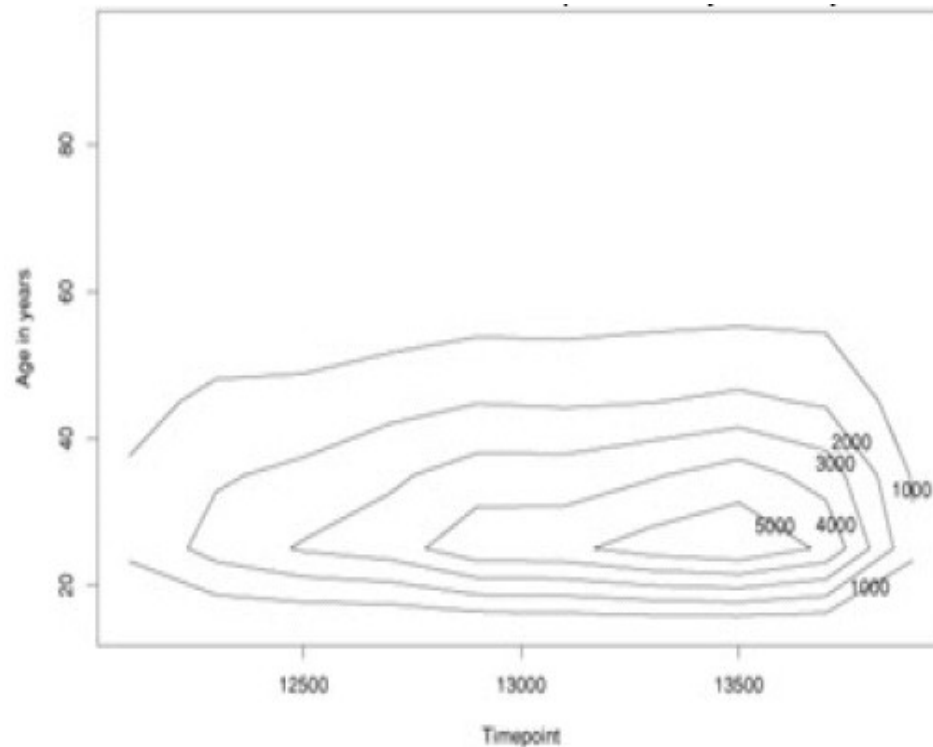
# 2-D: Saturación



- A veces se necesitan otros métodos

# 2-D: Contornos

- Apropriados para plots con alta densidad





# R: ejemplos

```
#dos dimensiones
```

```
x<-runif(100)*2*pi
```

```
y<-jitter(sin(x),amo=0.1)
```

```
y[c(1,100)]<-y[c(1,100)]*(-1)
```

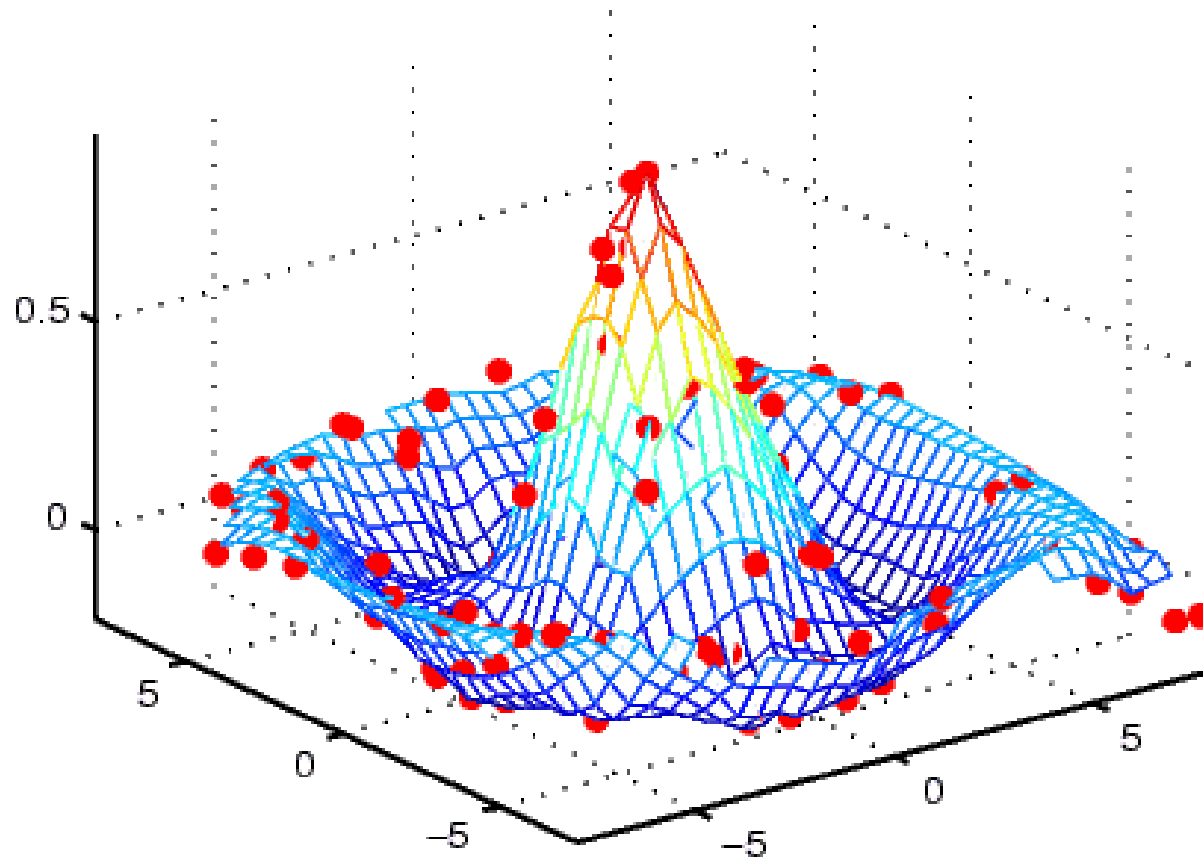
```
hist(x)
```

```
hist(y)
```

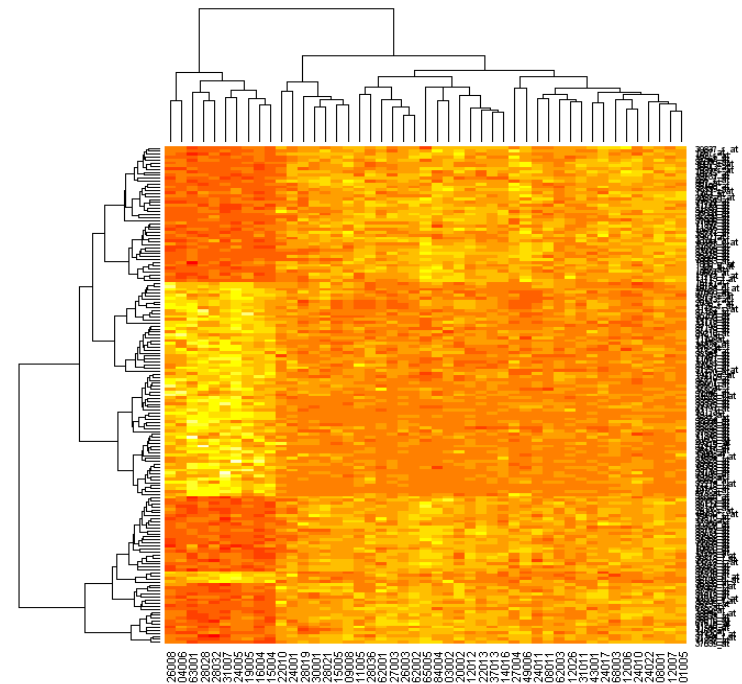
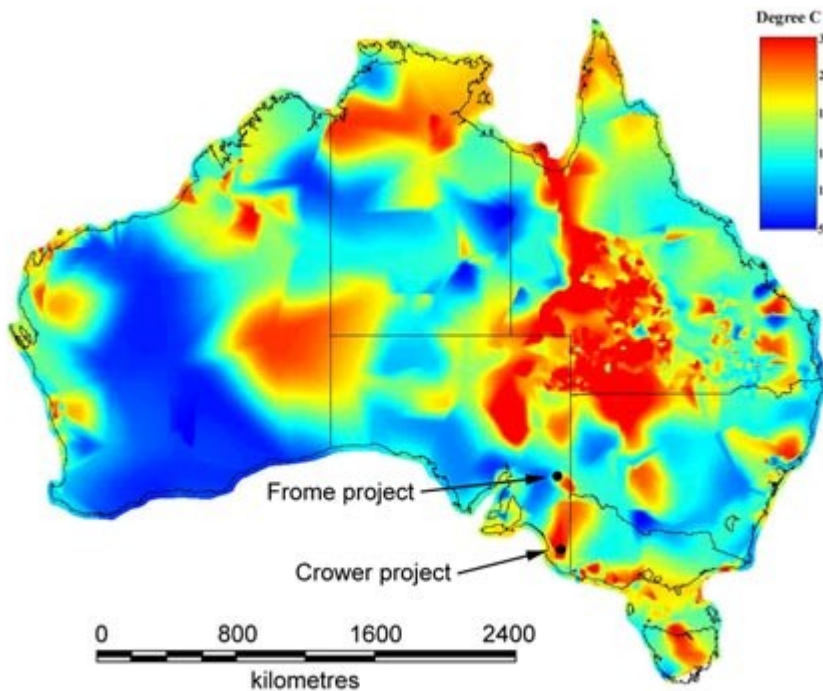
```
boxplot(y)
```

```
plot(x,y)
```

# Datos 3-D (proyecciones)



# Datos 3-D (heat-map)



# R: ejemplos

```
y<-x<-1:100*3.14/100  
z<-sin(x)%*%t(sin(y))  
image(x,y,z)  #heatmap  
library(graphics)  
contour(x,y,z)  
filled.contour(x,y,z)  
persp(x,y,z)
```

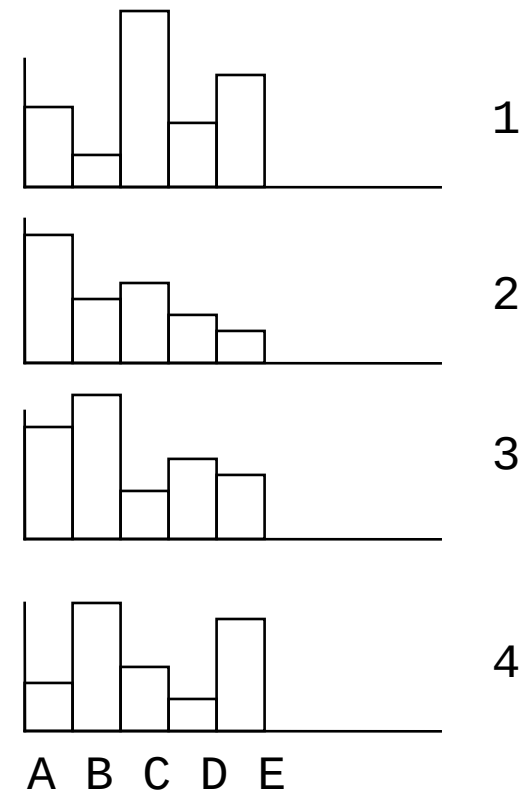
# Visualización en más dimensiones

- Scatterplots
- Parallel Coordinates
- Chernoff faces
- ...

# Vistas múltiples

Mostrar cada variable por separado

	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 muestra correlaciones

# Scatterplot

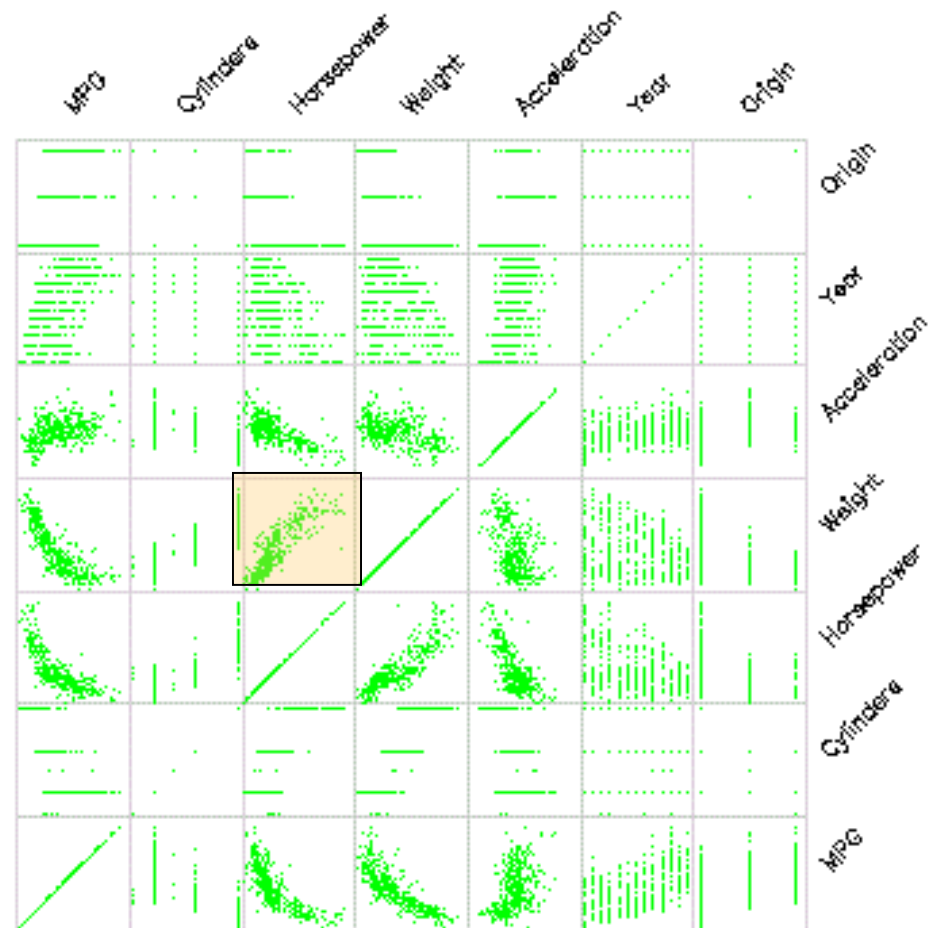
Muestra cada par de variables en un plot individual 2-D  
Ejemplo: car data

## *Ventaja:*

Se ven facilmente las correlaciones

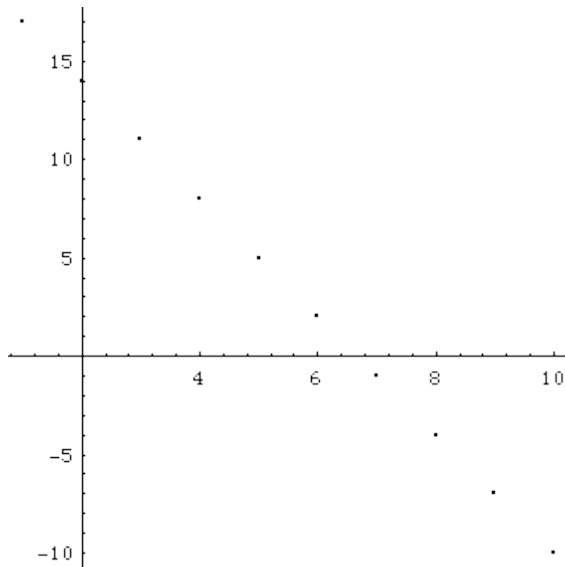
## *Problema:*

No se ven los efectos multivariados

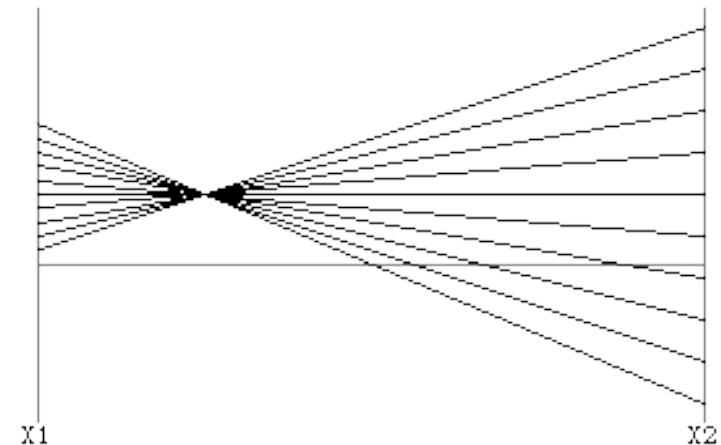


# Parallel Coordinates

- Pone cada variable en un valor distinto (fijo) del eje horizontal.
- Los valores de ponen en la vertical, y se unen con líneas



Un dataset en coordenadas  
Cartesianas




El mismo dataset en parallel  
coordinates



# Parallel Coordinates: ejemplo

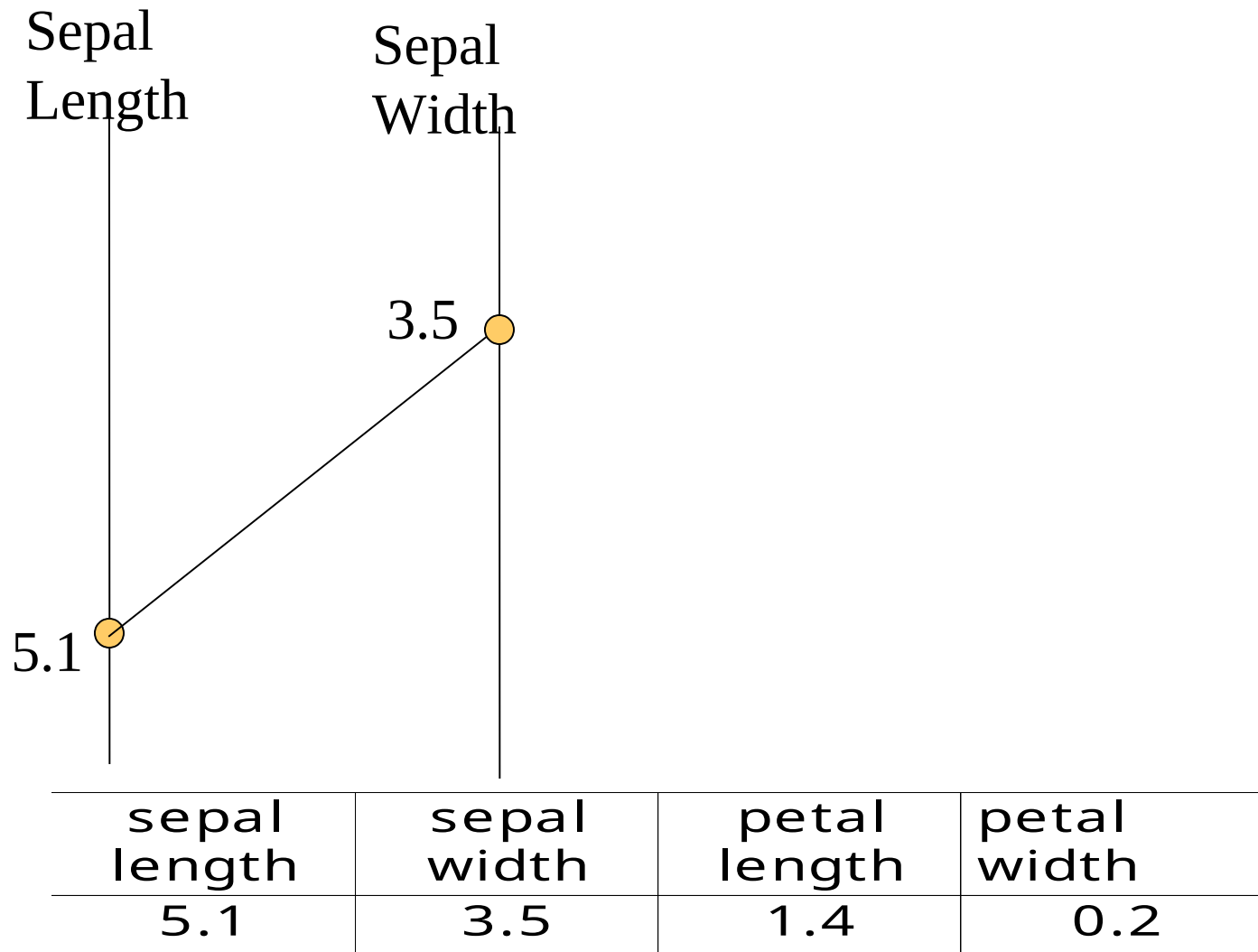
Sepal  
Length

5.1

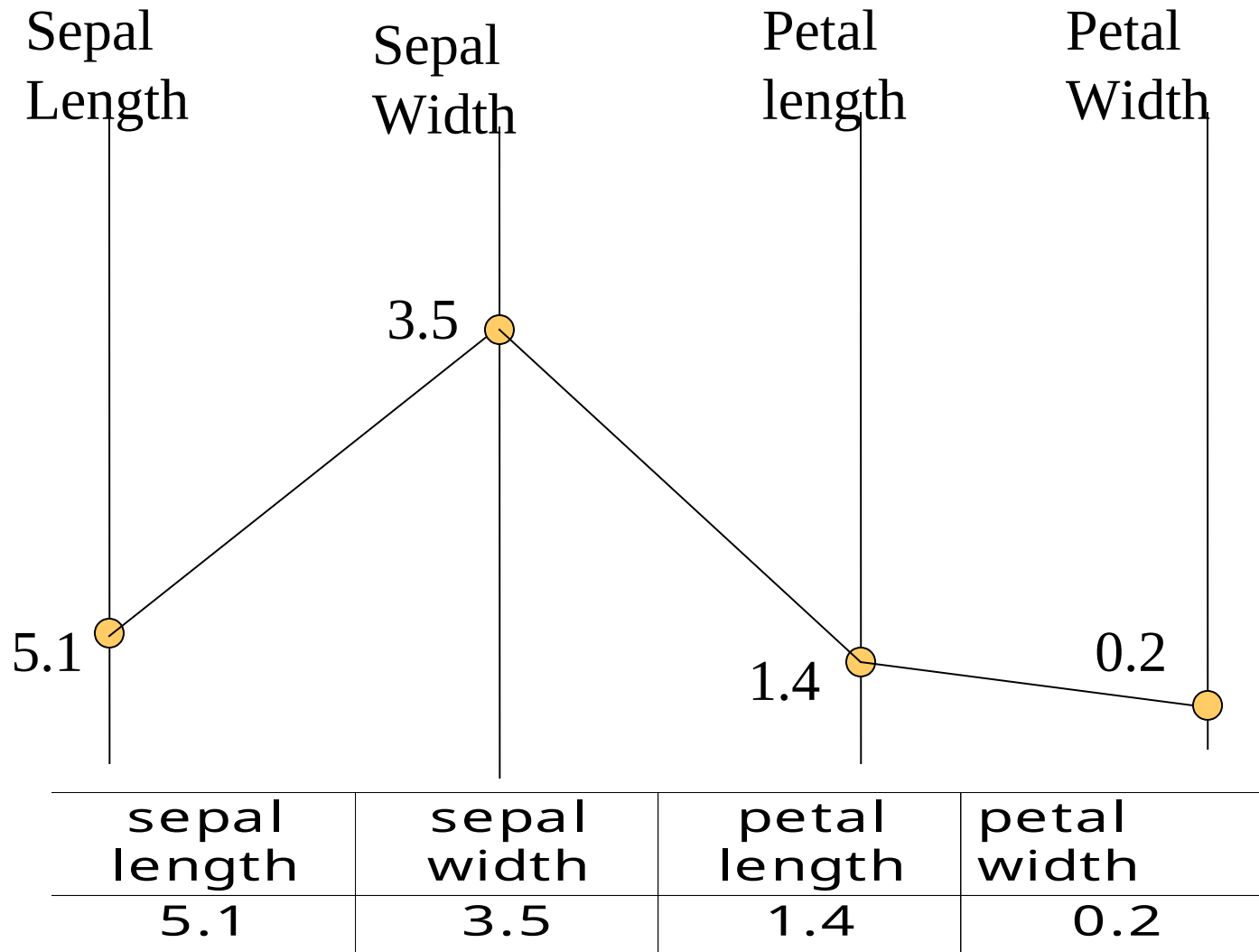


sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

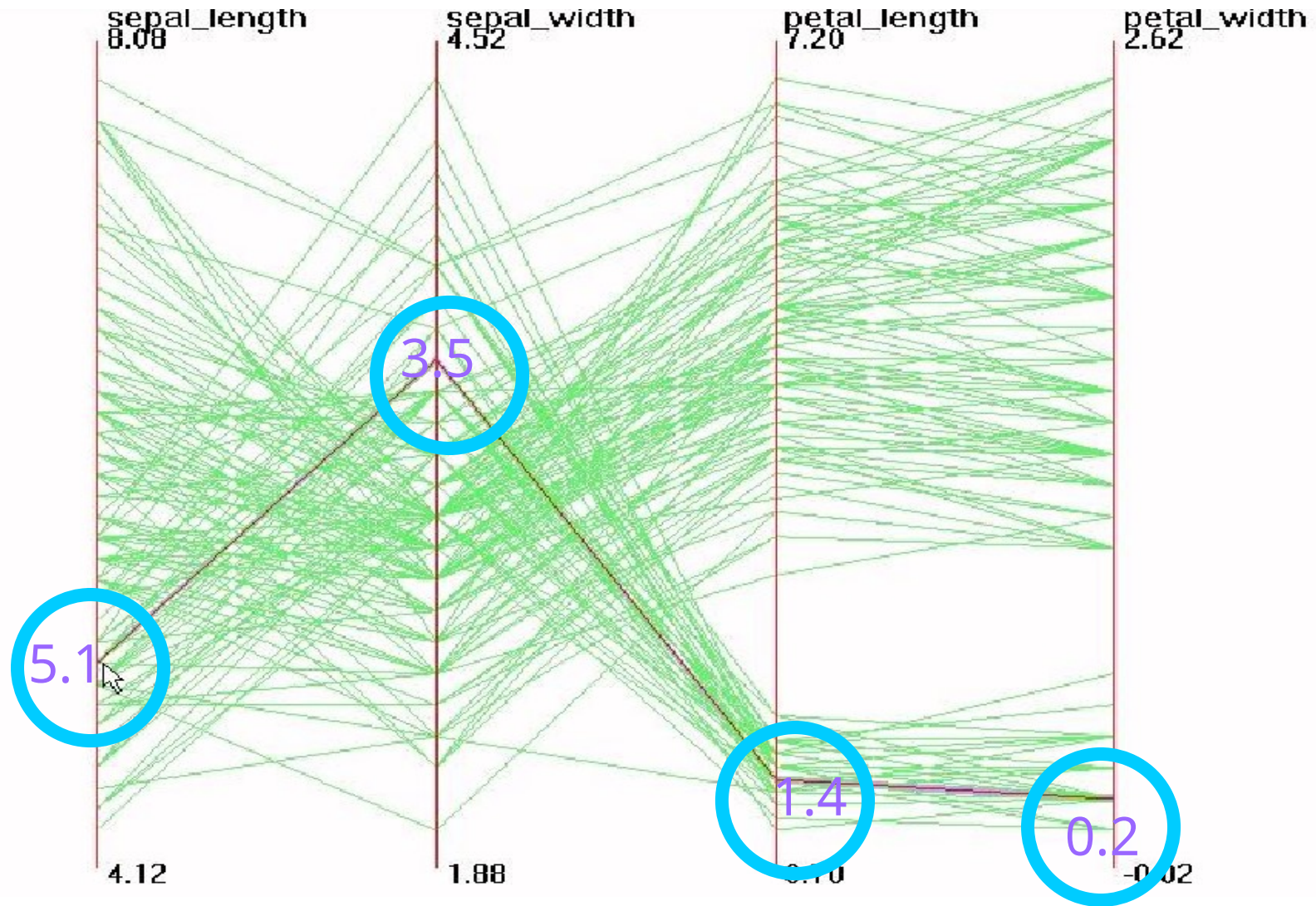
# Parallel Coordinates: 2 D



# Parallel Coordinates: 4 D



# Parallel Coordinates: Iris



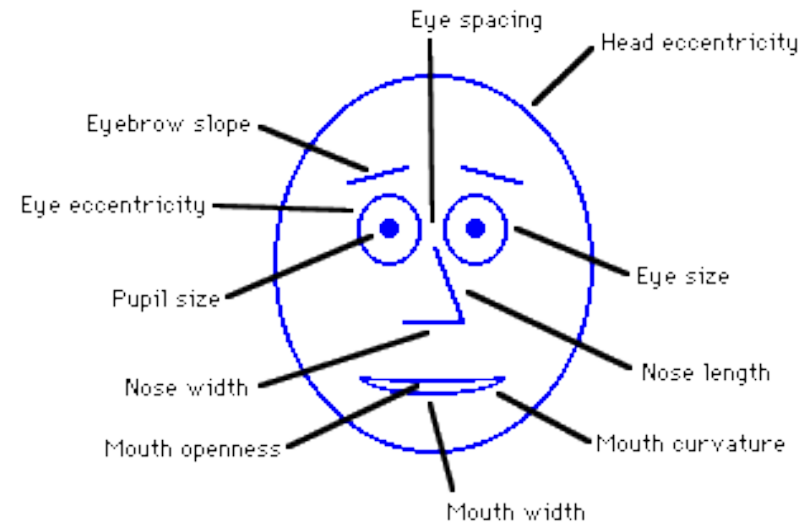
# Parallel coordinates: resumen

- Cada punto es una línea
- Puntos similares, líneas similares
- Las líneas que cruzan muestran atributos negativamente correlacionados
- Problemas:
  - el orden de los ejes es importante
  - Límite de ~20 dimensiones

# Chernoff Faces

Codifica las diferentes variables en características de la cara humana

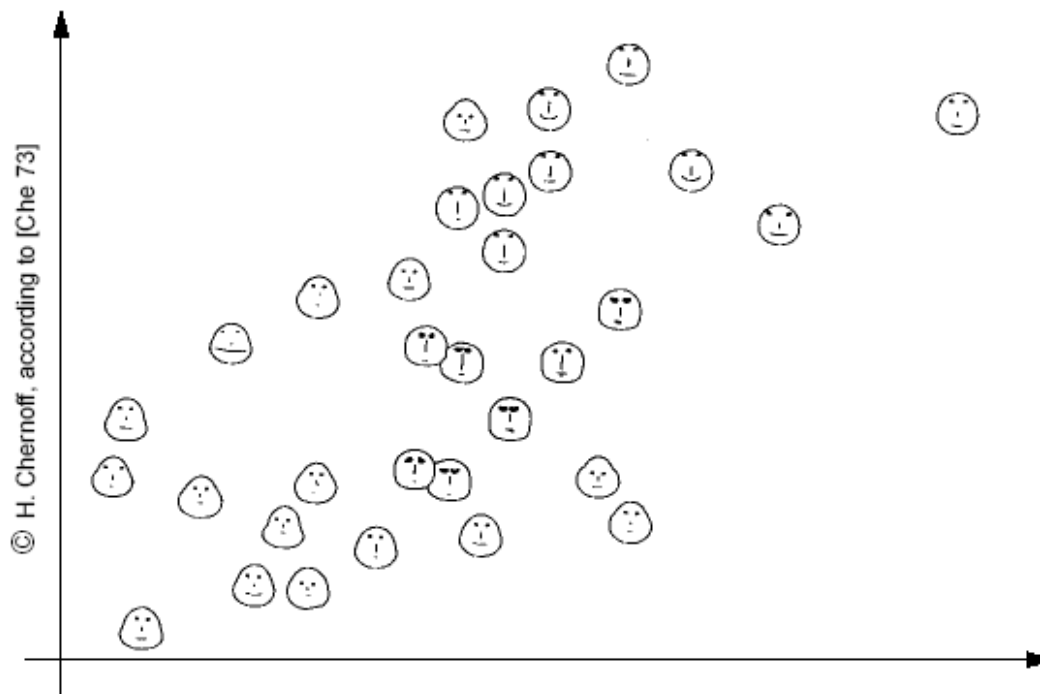
Aprovecha la capacidad humana de encontrar fácilmente pequeñas diferencias entre caras



Applets: <http://www.cs.uchicago.edu/~wiseman/chernoff/>  
<http://hesketh.com/schampeon/projects/Faces/chernoff.html>

# Chernoff faces, ejemplo

**Chernoff-Faces [Che 73, Tuf 83]**



# Stars plots

- Cada variable va en una dirección angular diferente. Cada punto forma una “estrella”



1



10



19



28



2



11



20



29





# R: ejemplos

```
data(iris)
```

```
summary(iris)
```

```
plot(iris[,-5],col=iris[,5])
```

```
library(denpro)
```

```
paracoor(iris[,-5],pal=iris[,5])
```

```
require(TeachingDemos)
```

```
faces(iris[,-5],ncol=25)
```

```
help(stars)
```

```
stars(iris[,-5],ncol=10,  
      col.sta=iris[,5])
```

```
boxplot(iris[iris[,5]== "setosa",-5])
```

```
boxplot(iris[iris[,5]==  
          "versicolor",-5])
```

```
boxplot(iris[iris[,5]== "virginica",-  
            5])
```

# Resumen

- Muchos métodos, distintas ventajas
- Se pueden visualizar datos en más de 3-D
- Buscar siempre:
  - Hacer buenas gráficas
  - Que muestren la mayor cantidad de información
  - Que no mientan sobre los datos