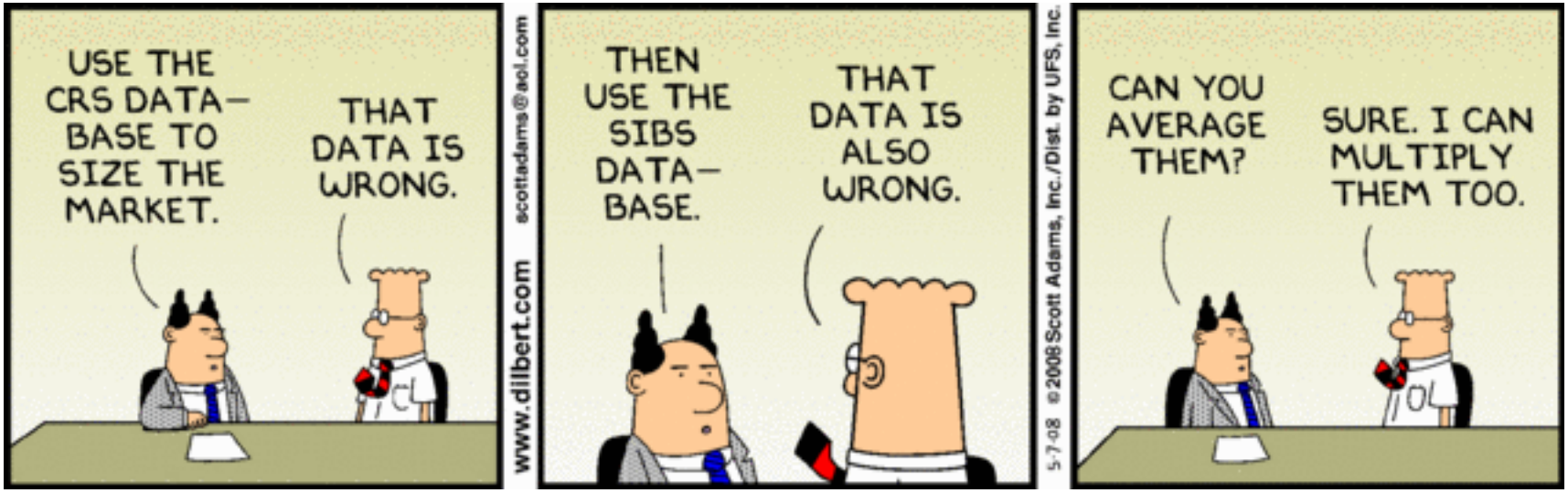
Data Preprocessing

Preparación de la información



Data Preprocessing

¿Por qué preprocesar los datos?

- **Datos incompletos:** falta de valores en algunos atributos, datos que vienen sólo agregados.

- **Datos ruidosos:** Errores de ingreso, outliers

Data Preprocessing

¿Por qué preprocesar los datos?

- **Datos inconsistentes:** Diferencias en nombres de atributos para distintas áreas de la compañía, diferencias de unidades, codificaciones, mismo registro con distintos nombres en distintas bases de datos, etc.

- **Muchos datos:** A veces es necesario reducir la información para hacer el análisis

Análisis descriptivo de los datos

"Data don't make any sense,
we will have to resort to statistics."

Análisis descriptivo de los datos

- Objetivo: Tener una visión de algunas características generales de los datos

- Es útil para el análisis inicial de la información.

- Ejemplo de algunas medidas descriptivas iniciales: media, mediana, varianza, etc.

Medidas de tendencia Central

- Media aritmética: $$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Media aritmética con pesos (weighted aritmetic mean):

$$\overline{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

- Algunos problemas: Sensibilidad a valores extremos

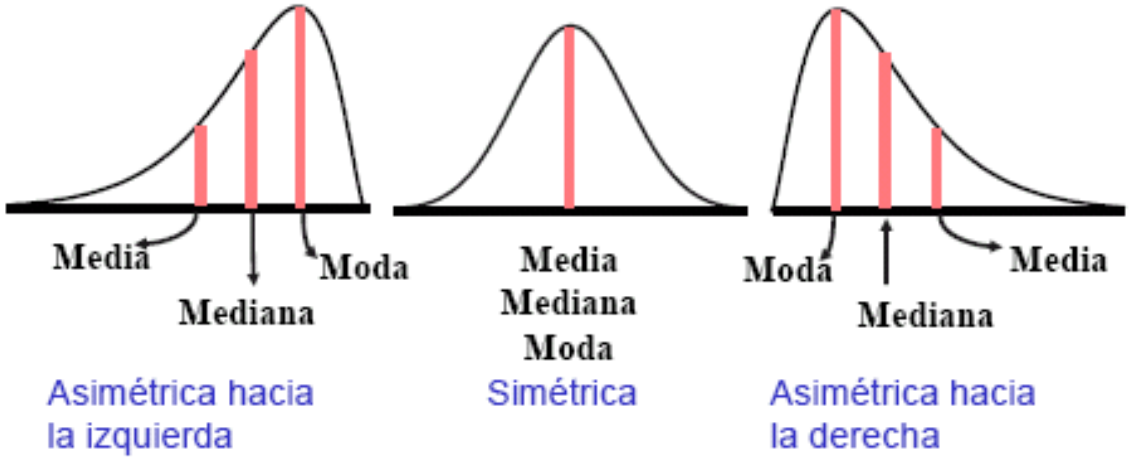
Medidas de tendencia Central

- **Media recortada (trimmed mean):** Se excluyen los valores más extremos para el cálculo de la media, ej. Se puede eliminar el 2% más alto y el 2% más bajo de los datos. Valores muy altos de exclusión causan pérdida de información

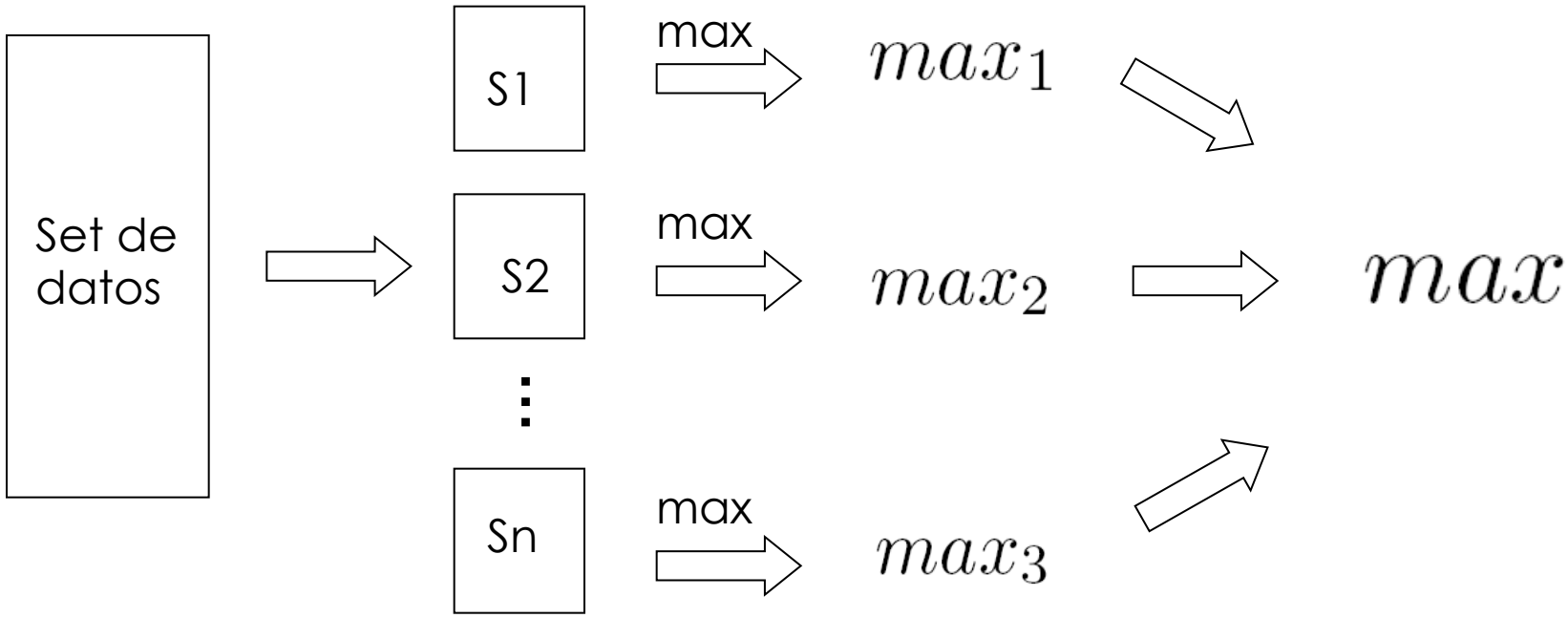
- **Mediana:** Utilizada más en datos asimétricos. Si ordenamos los números en un arreglo de tamaño N, la mediana corresponde al valor central del arreglo si N es impar, y al promedio de los dos valores del centro si N es par.

Medidas de tendencia Central

- **Moda:** Es el valor que tiene más frecuencia en el set de datos.



- **Medidas distributivas:** Medidas que se pueden obtener computando subconjuntos de los datos y luego mezclando los resultados (ej. sum,count,max,min)



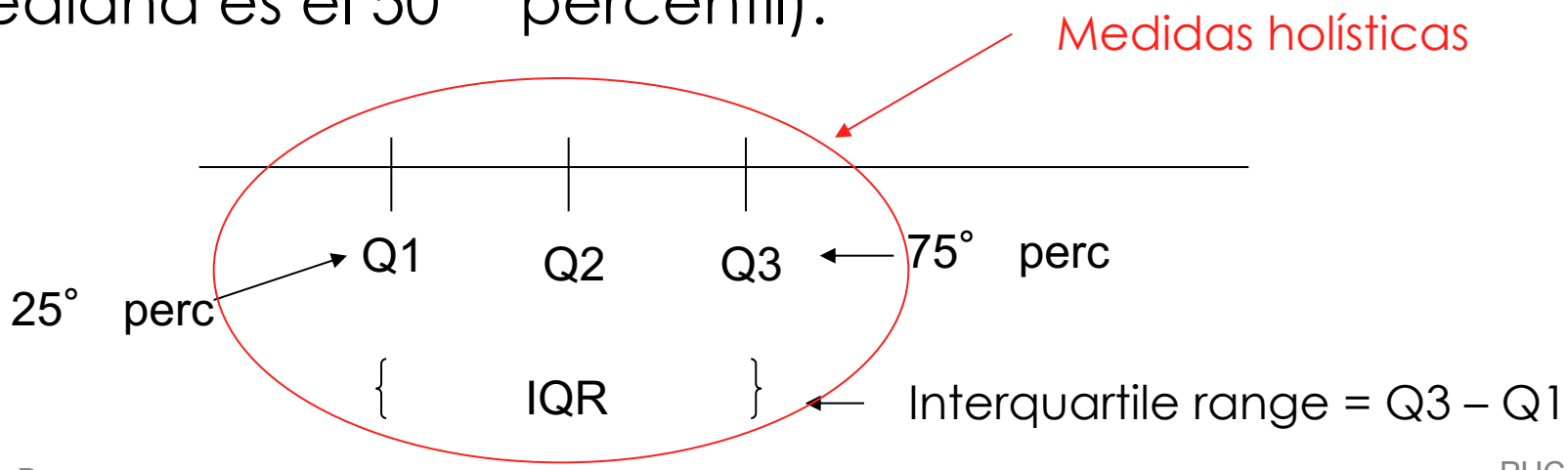
- Medida holística: Medidas que sólo se pueden obtener computando el set de datos completo como un todo. Ej Mediana, promedio*.

Las medidas holísticas son más caras de computar

Medidas de Dispersión de los datos

- **Rango:** Para un set de datos $x_1, x_2, \ldots, x_N$ (Observaciones de un atributo), corresponde a la Diferencia entre el mayor y el menor valor

- **K-ésimo percentil:** Para un set de observaciones ordenadas en forma creciente corresponde al valor para el cual el K% de los datos queda antes que él (la mediana es el 50° percentil).

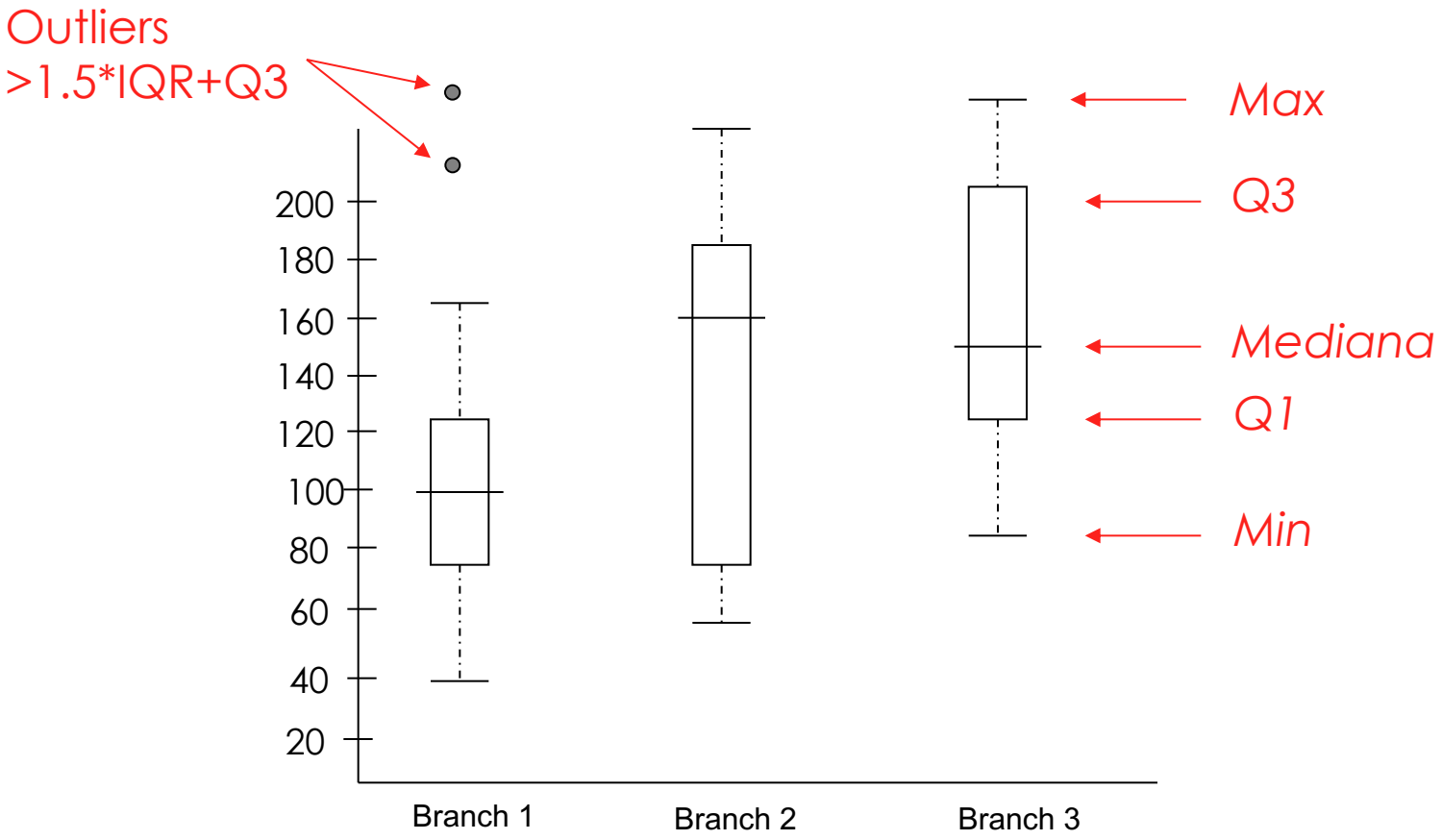


100 - quantiles = percentiles
10 - quantiles = deciles
5 – quantiles = quintiles
4 – quantiles = cuartiles

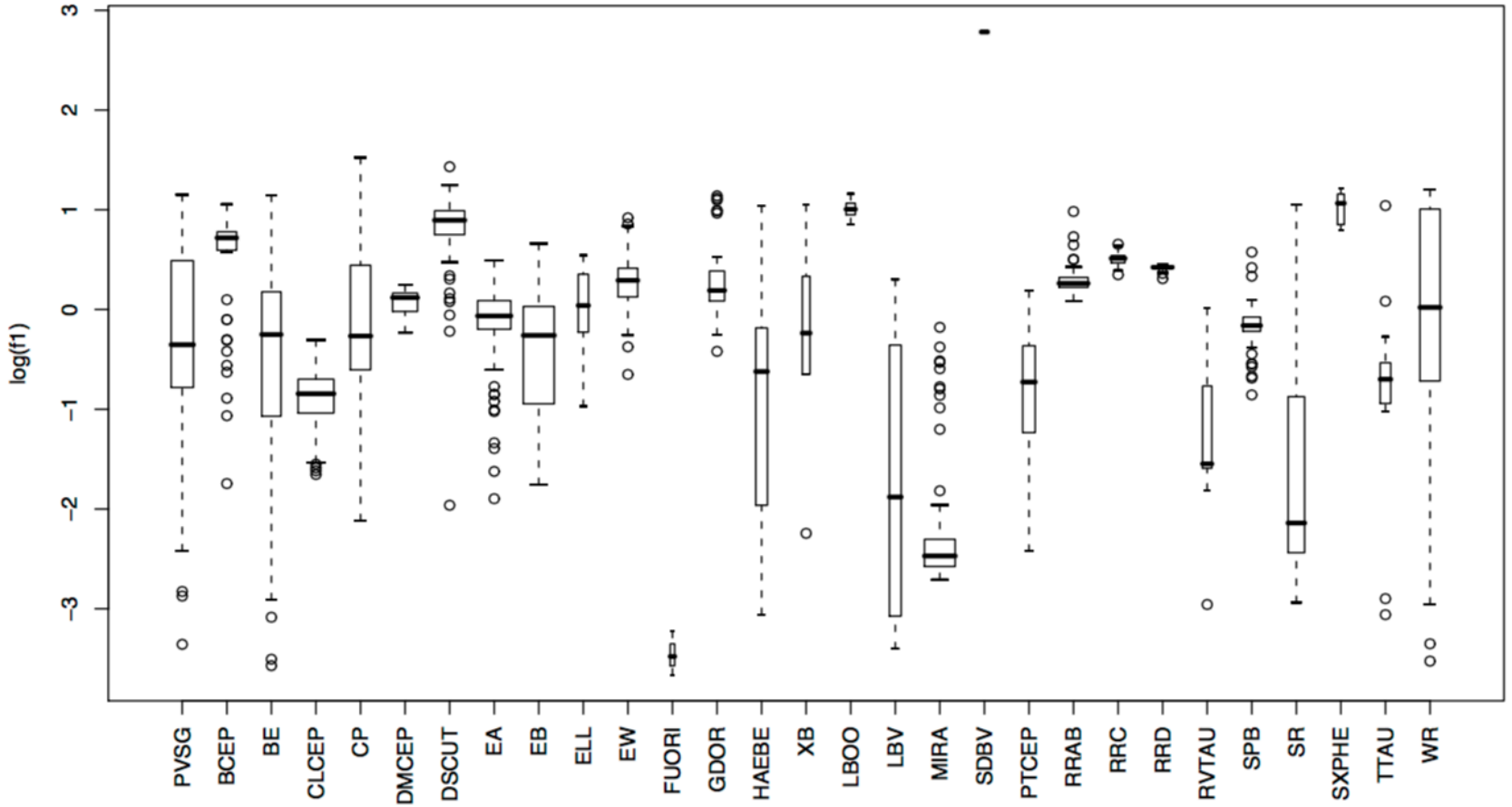
Ej. De uso de IQR: En detección de outliers, para valores a una distancia mayor a 1.5*IQR por sobre Q3 o bajo Q1.

- ***Five number summary:*** Corresponde a la secuencia de los valores *Min, Q1, Mediana, Q3, Max*

Boxplots: Forma gráfica de visualizar estos 5 valores:



Ex: Boxplots per class for one variable



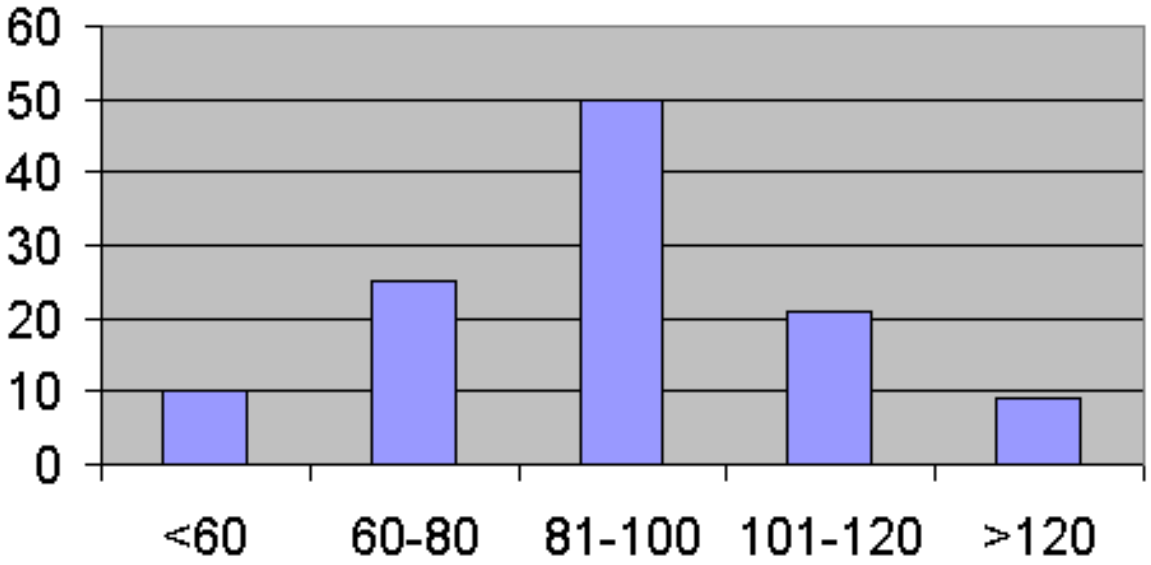
Fuente: https://arxiv.org/pdf/0711.0703v1.pdf

Boxplots in Python

- Example in python:

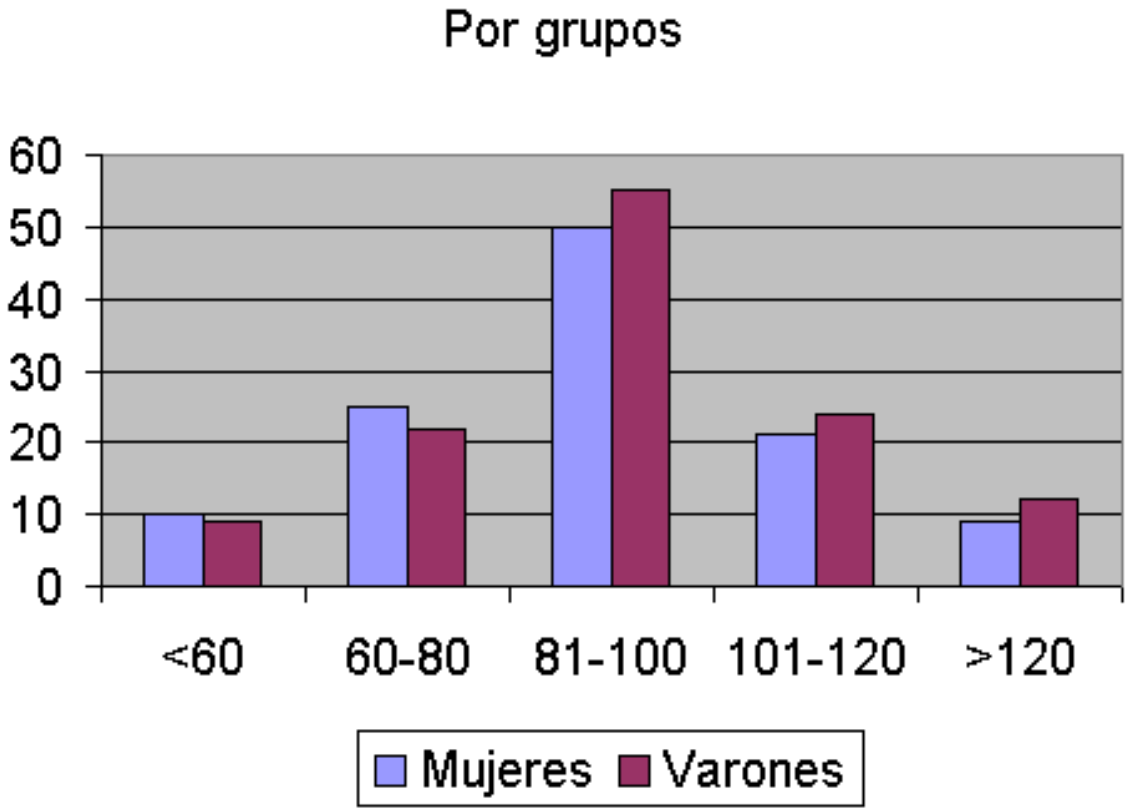
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

#create some artificial data
s = pd.Series((5, 15, 10, 15, 5, 10, 10, 20, 25, 15))
plt.boxplot(s)
plt.yticks(range(max(s)))
plt.show()
```

Algunos Gráficos

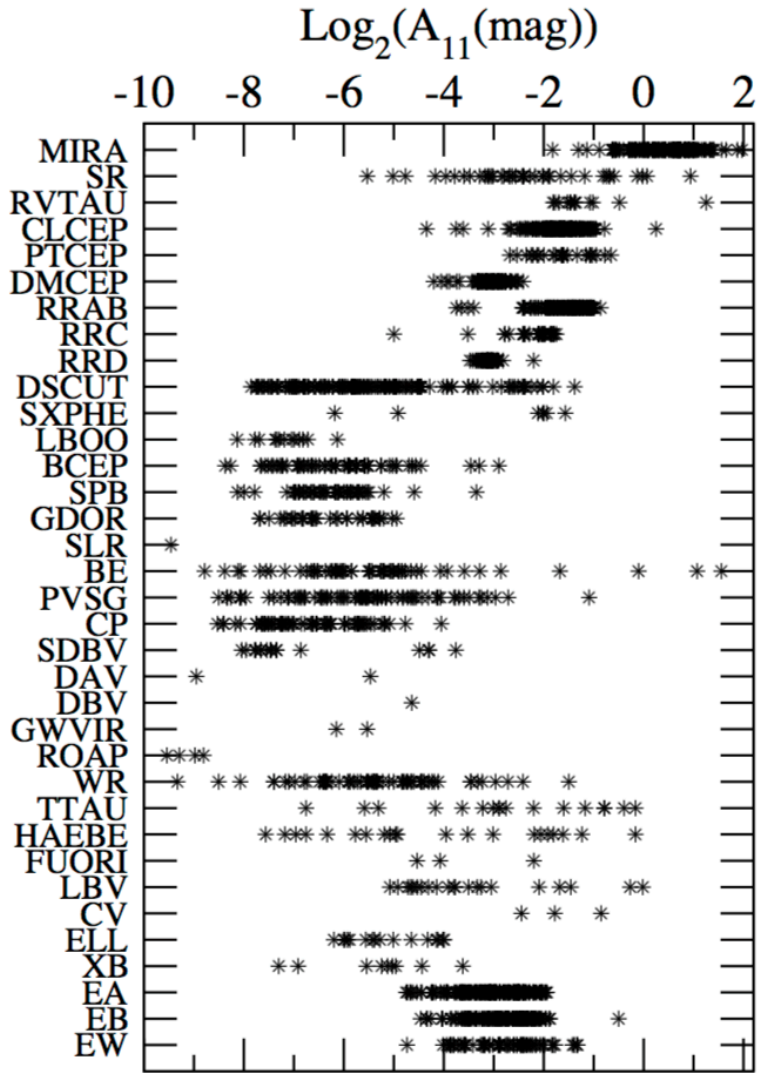


Histograma Simple

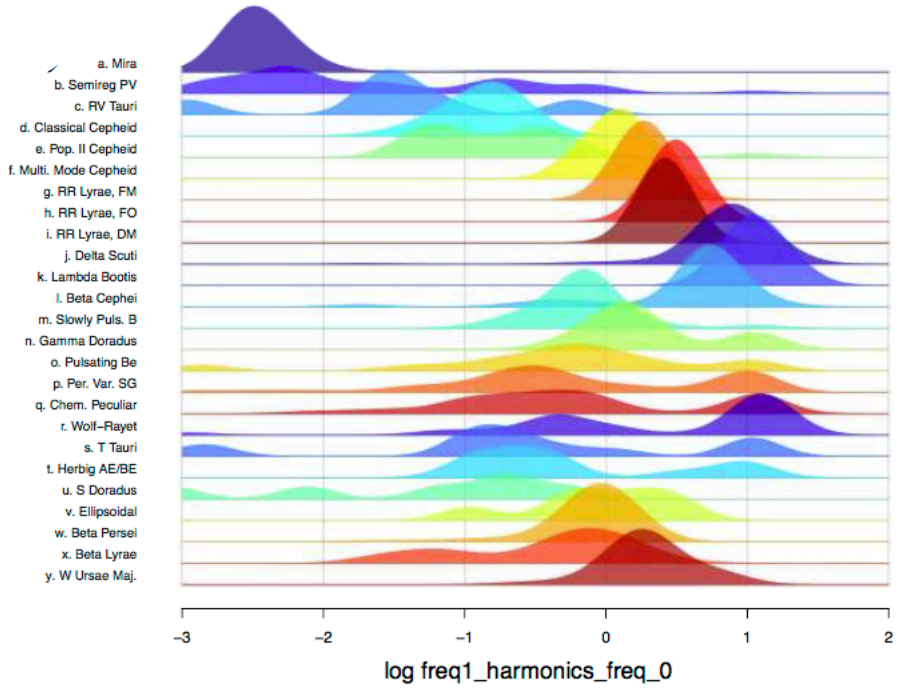
Algunos Gráficos(cont..)



Ex: Data distribution per class for one variable

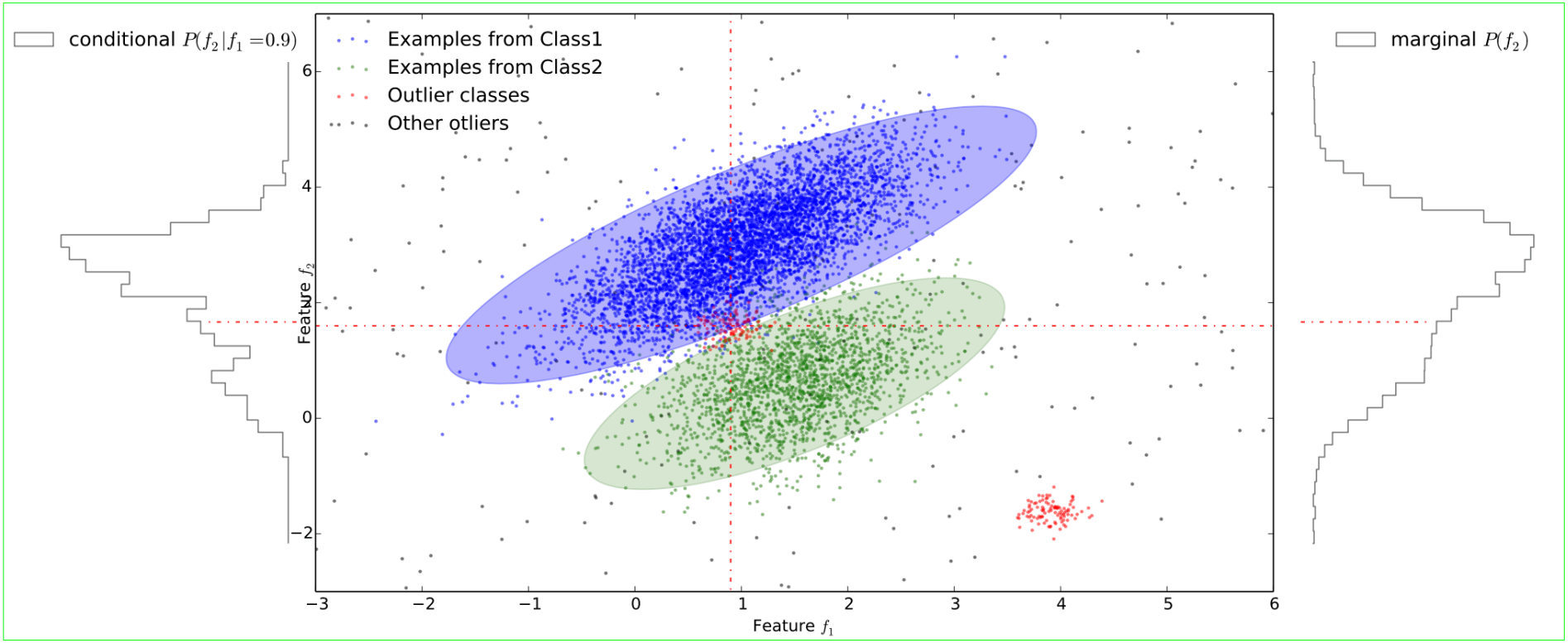


https://arxiv.org/pdf/0711.0703v1.pdf



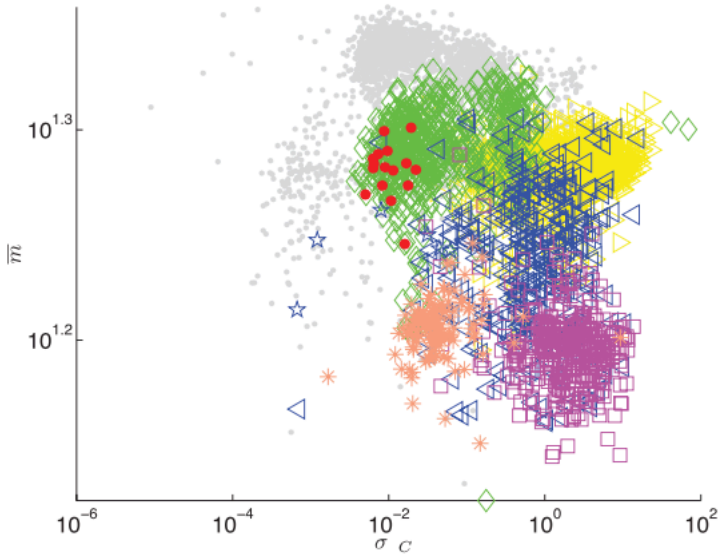
http://arxiv.org/pdf/1104.3142.pdf

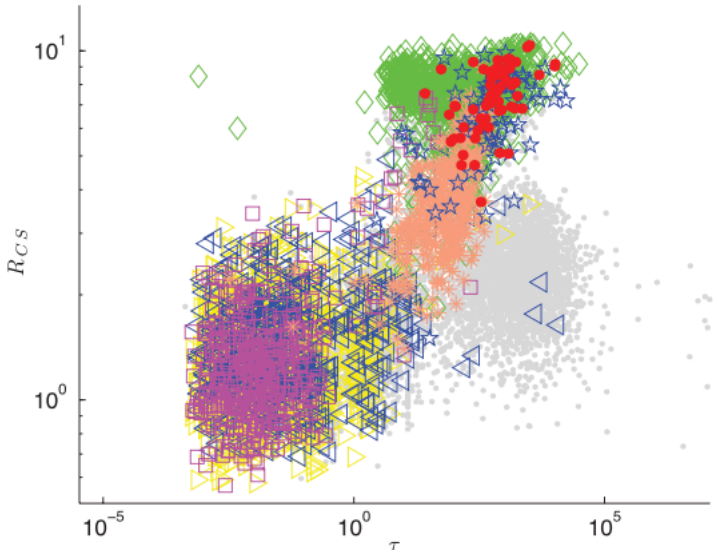
Ex: Histograms for conditionals and marginals

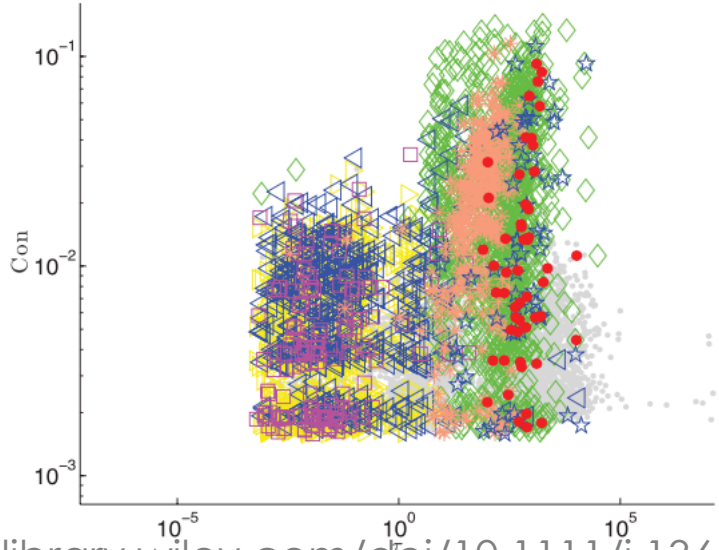


http://arxiv.org/pdf/1404.4888.pdf

Ex: Data distribution per class for two variables



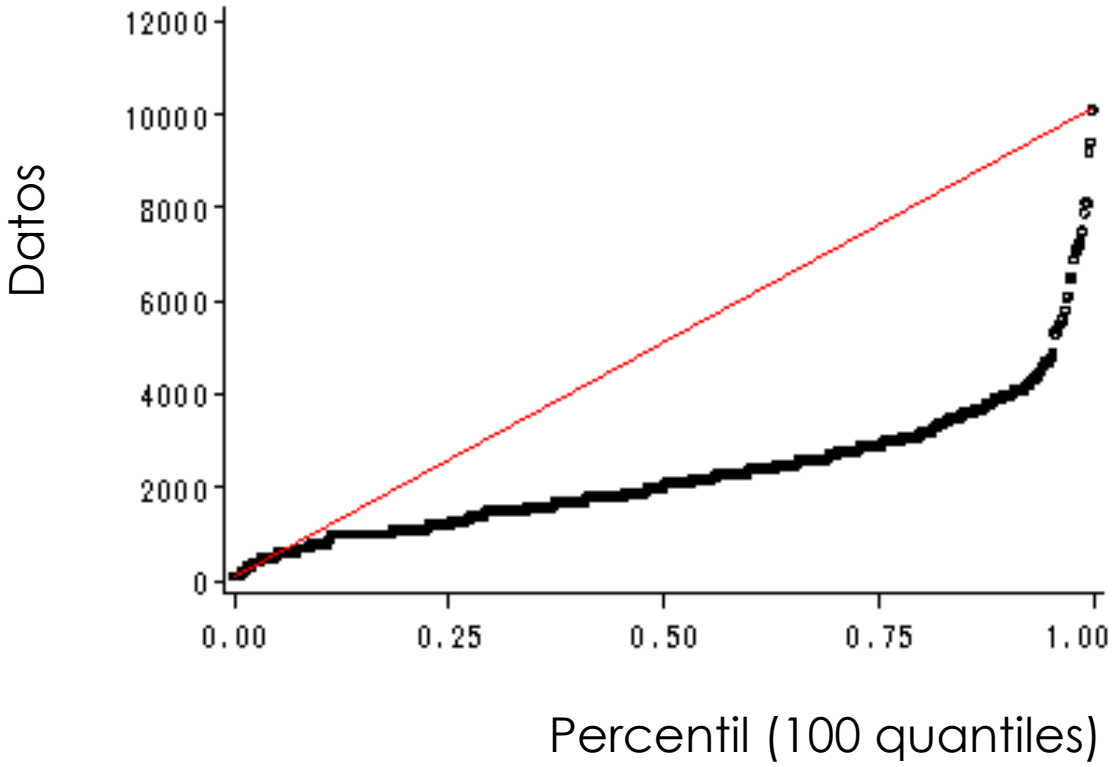




Fuente: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2966.2012.22061.x/epdf

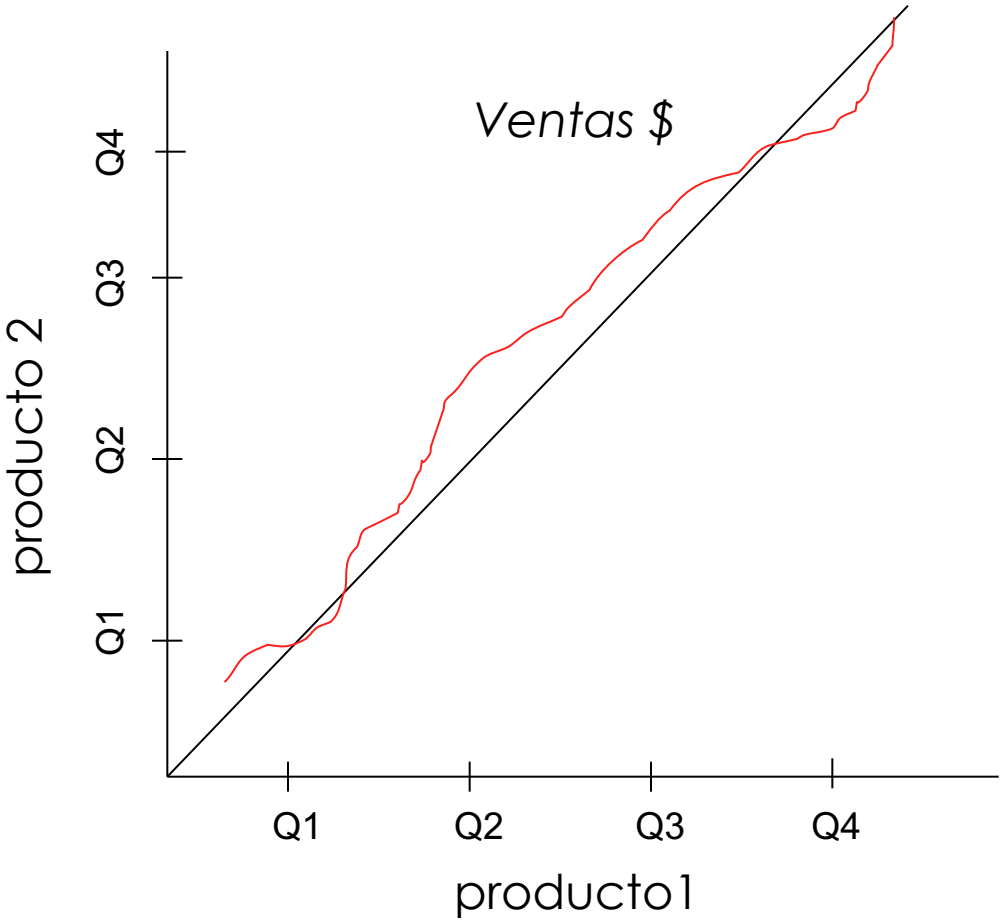
Algunos Gráficos(cont..)

Quantile Plot



Algunos Gráficos(cont..)

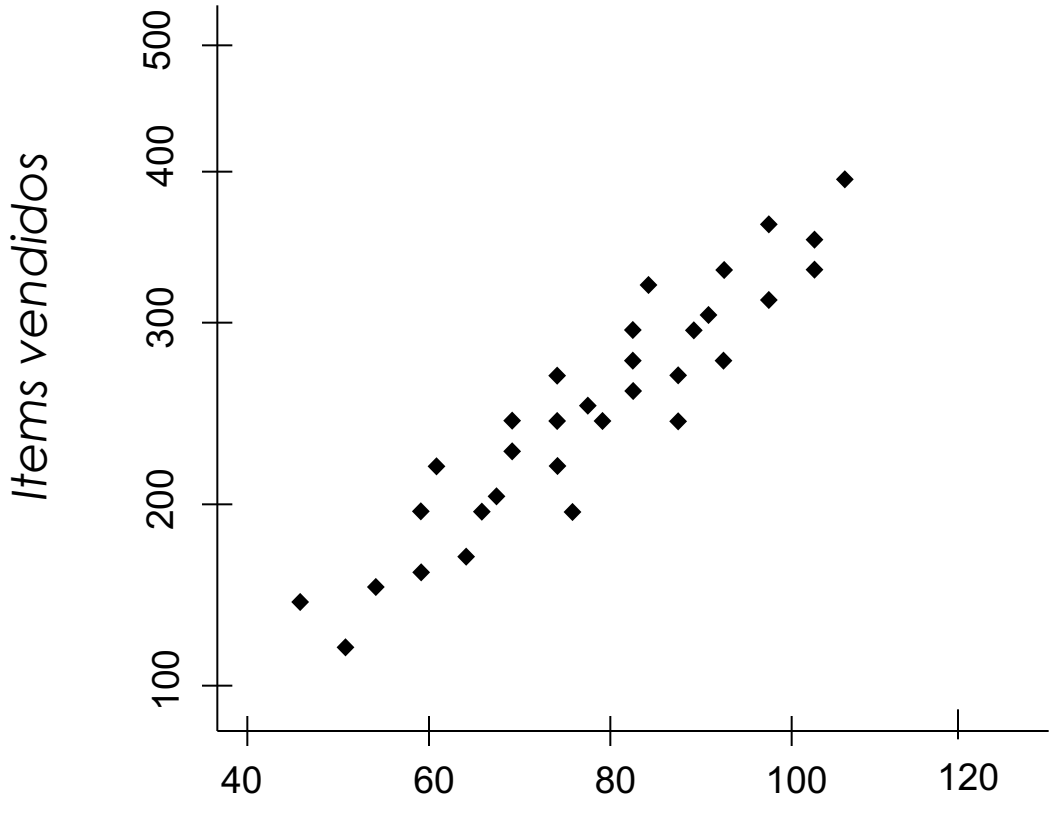
Quantile-Quantile Plot



Ventas tienden a ser menores en producto 1.

Algunos Gráficos(cont..)

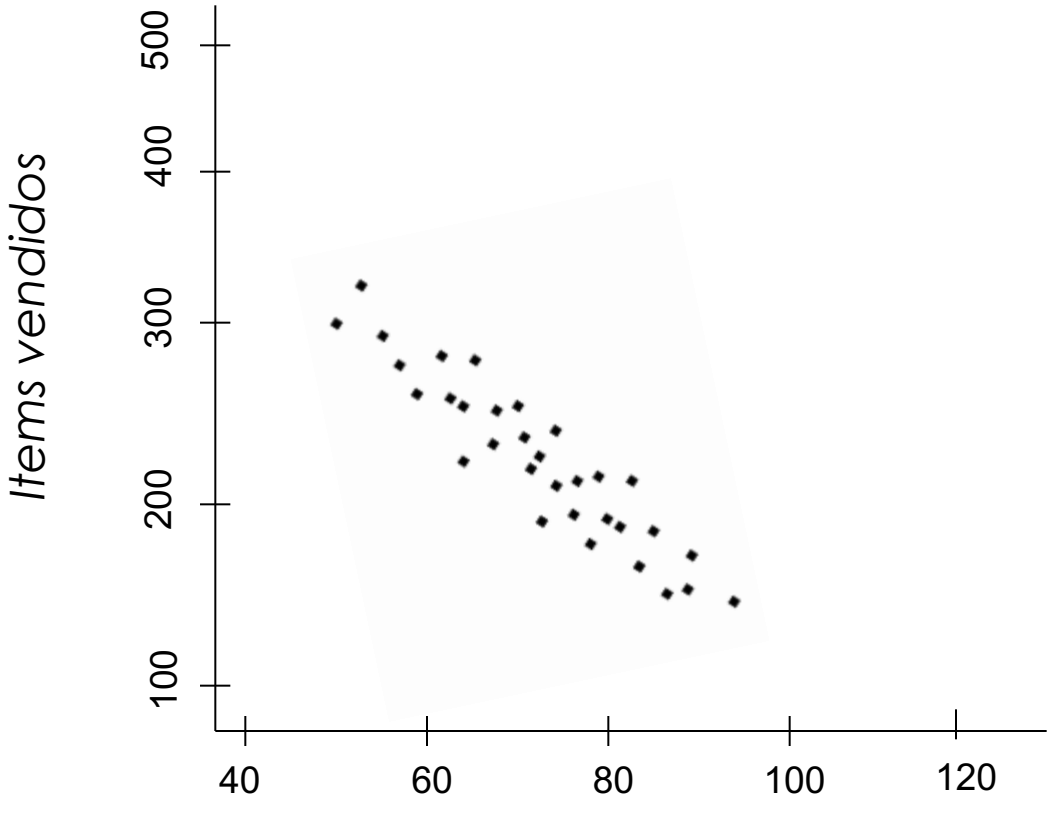
Scatter Plot



correlación positiva

Algunos Gráficos(cont..)

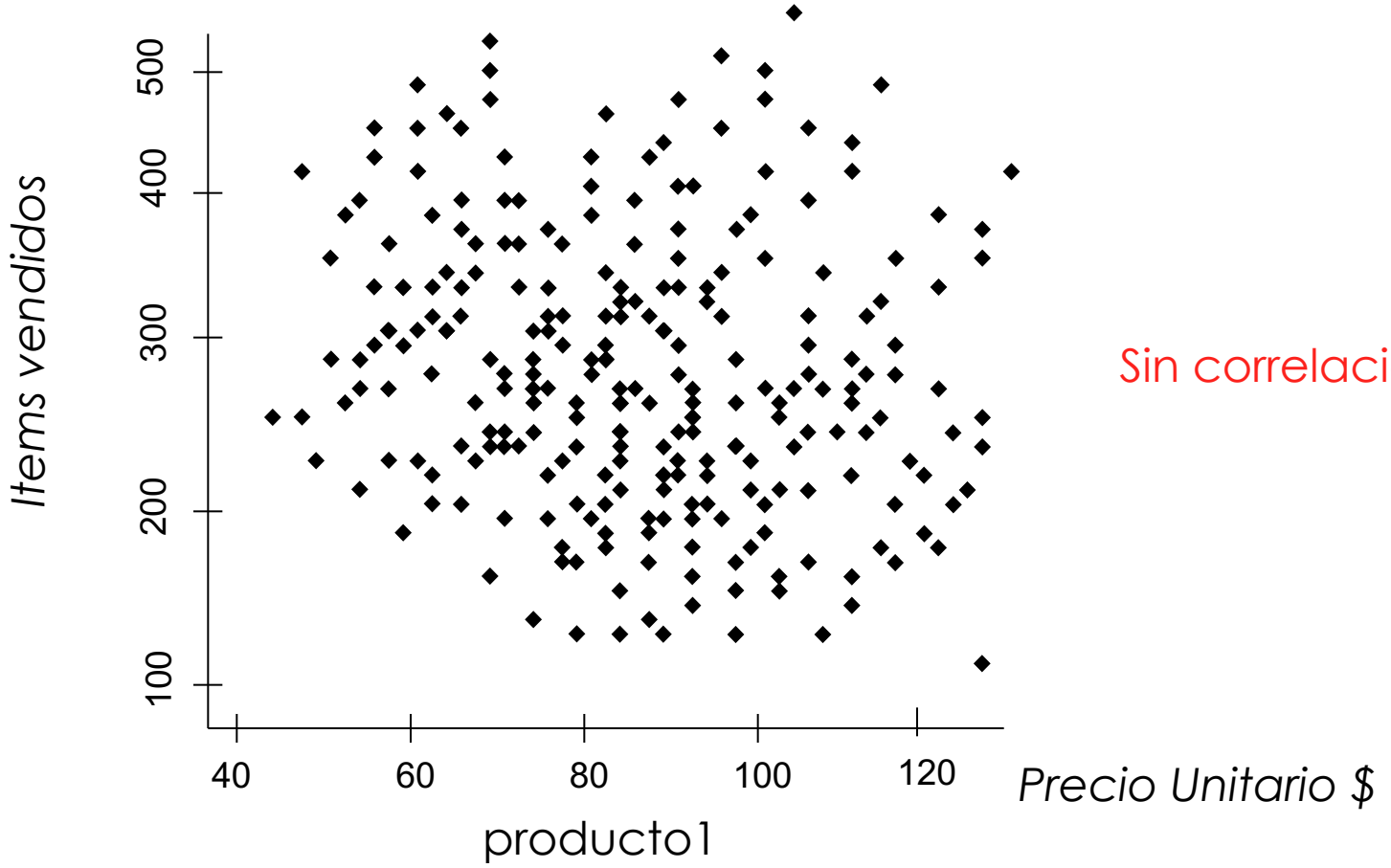
Scatter Plot



correlación negativa

Algunos Gráficos(cont..)

Scatter Plot



Sin correlación

Data Cleaning

- Datos en el mundo real tienden a ser incompletos, ruidosos e inconsistentes.

- El proceso de limpieza de datos trata de llenar los valores que faltan, identifica valores erróneos tratando de corregirlos y elimina inconsistencias en la información.

Datos Faltantes

- Muchas veces un atributo viene vacío
- Esta situación afecta el proceso de análisis
- Existen varias opciones para solucionar el problema:

Datos Faltantes(Cont..)

- **Ignorar la tupla:** Método poco efectivo a menos que falten muchos atributos en la misma fila.
  Problemas cuando faltan valores en pocos atributos (aleatoriamente) pero en muchas tuplas .

- **Llenar los valores manualmente:** No es practicable cuando el set de datos presenta muchos valores faltantes

- **Usar una cte. Global para llenar los valores:** Ej: "desconocido", "-∞",etc. Trae problemas para algunos algoritmos de data mining que considerarían estos valores como datos válidos y trataría de encontrar patrones para ellos

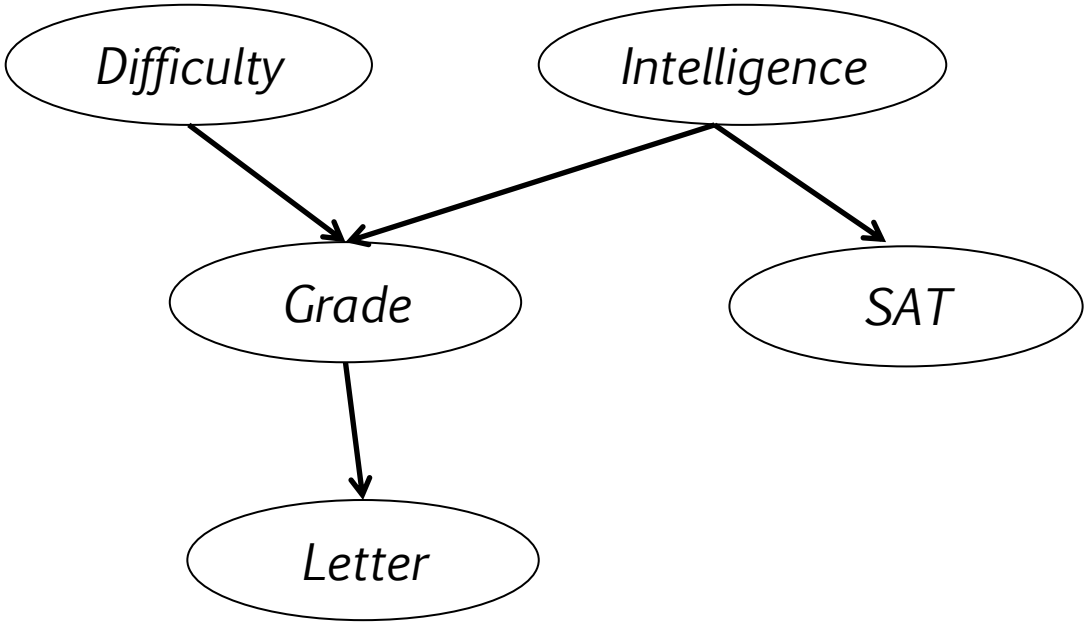
Datos Faltantes(Cont..)

- **Usar la media del atributo:** Llenar todos los valores faltantes en dicho atributo con el valor de la media para ese atributo. Es poco exacto

- **Usar la media por clases:** Igual que el método anterior pero utilizando la media considerando sólo los elementos que corresponden a la misma clase. Ej: Si falta el valor correspondiente al sueldo de un cliente de la clase business, llenarlo con el promedio del sueldo de todos los clientes business.

- **Usar el valor más probable:** Este valor puede ser determinado por regresión, herramientas de inferencia, árboles de decisión, etc.

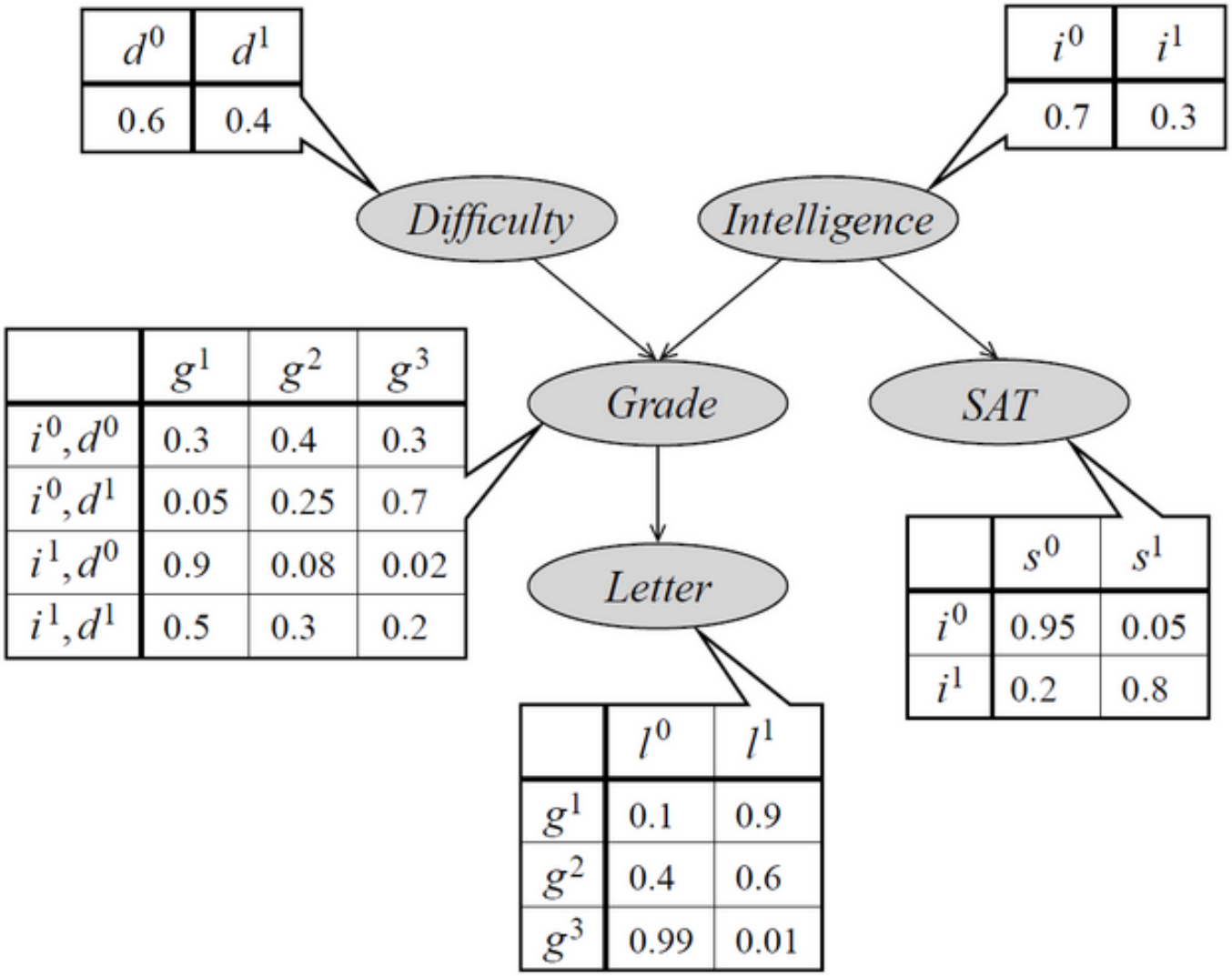
Ex Valor más probable: Bayesian Networks



BN Factorization:

$$P(D,I,G,S,L) = P(D)P(I)P(G \mid D,I)P(S \mid I)P(L \mid G)$$

Ex: Valor más probable



d^0	d^1
0.6	0.4

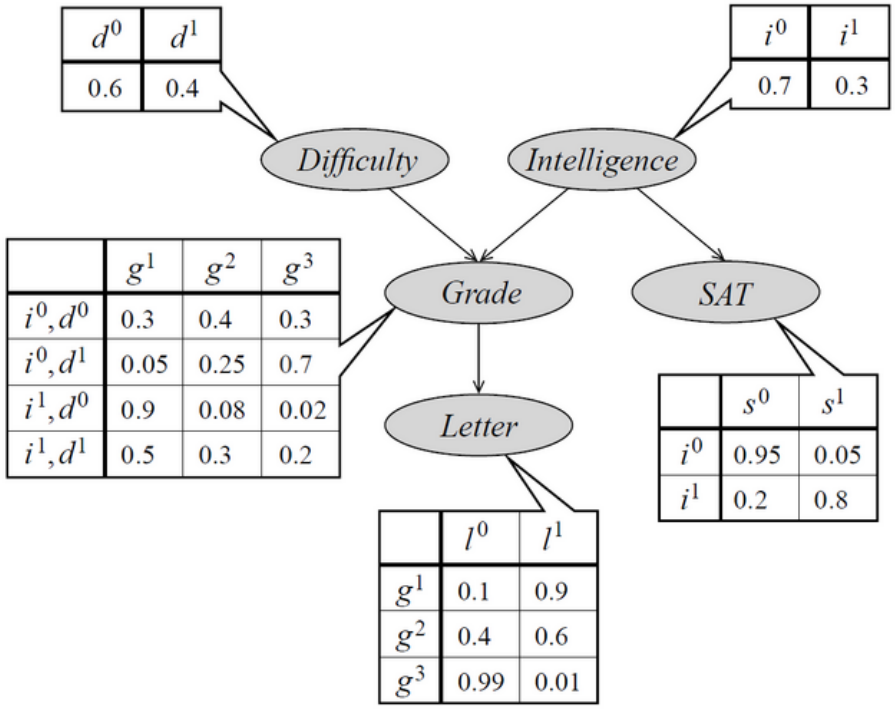
i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Ex: Valor más probable



Ej: Qué hacemos si Grade es faltante, por ejemplo, tenemos una entrada:

x = [0 , 1 , ? , 0, 0] (D, I, G, S, L) respectivamente

Datos Faltantes(Cont..)

- No siempre un dato faltante es un error. Ej, persona no tiene licencia de conducir, no usa tarjeta de crédito, etc.
- En esos casos es importante tener valores definidos como “no se aplica”, etc.

Algunas técnicas de preprocesamiento

- **Binning:** Los datos se ordenan separándose en grupos (bins).
  - Smoothing by bin means: Cada valor en el bin es reemplazado por la media del bin.
  - Smoothing by bin boundaries: cada valor se reempla za por el valor mínimo del bin o el máximo dependiendo de cuál sea el más cercano.

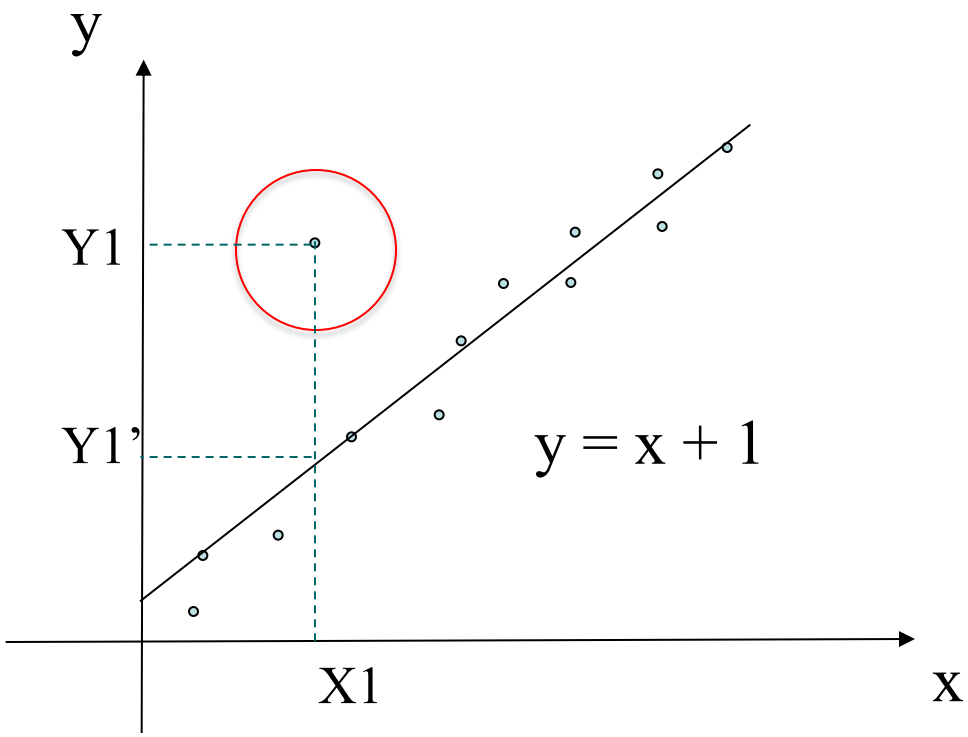
Este método se utiliza como herramienta de Discretización, smoothing, etc.

* Sorted data : 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Algunas técnicas de preprocesamiento

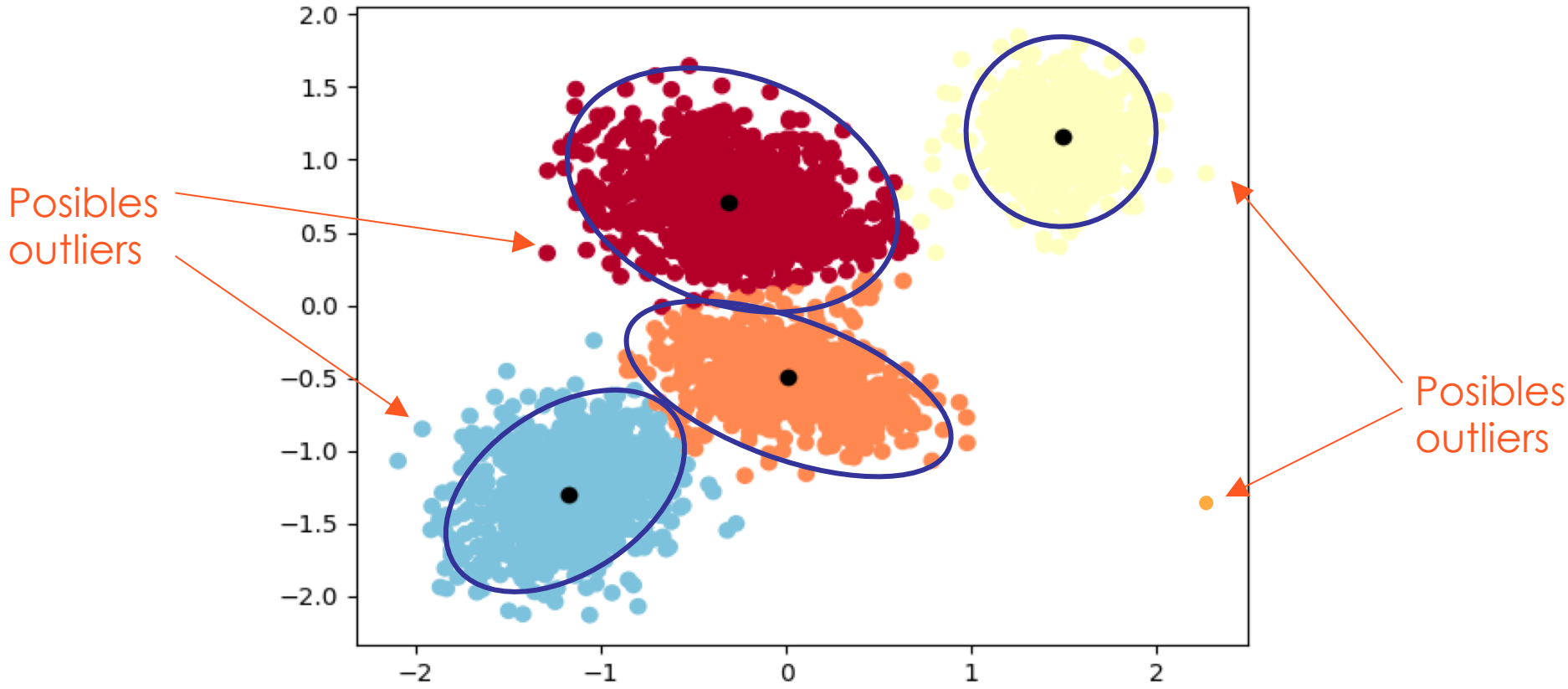
- Regresión para corrección: Algunos datos se "corrigen" en base a una función. Ej:

Regresión Lineal



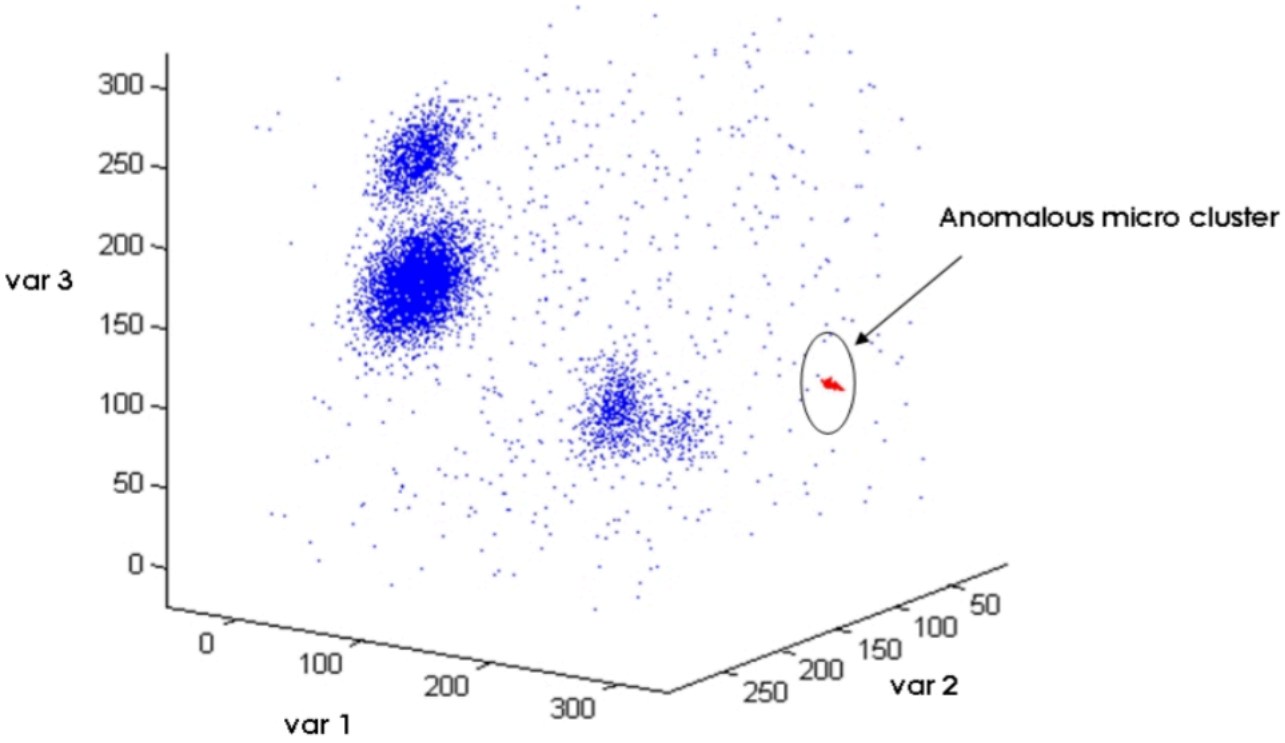
Algunas técnicas de preprocesamiento

- **Clustering** para la detección de outliers (candidatos a ser datos erróneos)



Algunas técnicas de preprocesamiento

- **Clustering** para la detección de outliers: Outliers are not always alone



Integración

- La información en la mayoría de los casos debe ser integrada desde múltiples fuentes de datos.

- Algunos problemas típicos:
  - *Identificación de la entidad*
  - *Redundancia*
  - *Detección y Resolución de conflictos entre valores*

Identificación de la entidad (entity identification problem)

- La misma entidad tiene distintos nombres en diferentes fuentes de datos, ej, customer_id, cust_number.

- Para esto se utilizad Metadata donde se almacena información sobre las entidades en cada fuente de datos, ej: nombre, significado, tipo de datos, rango, valores nulos, etc.

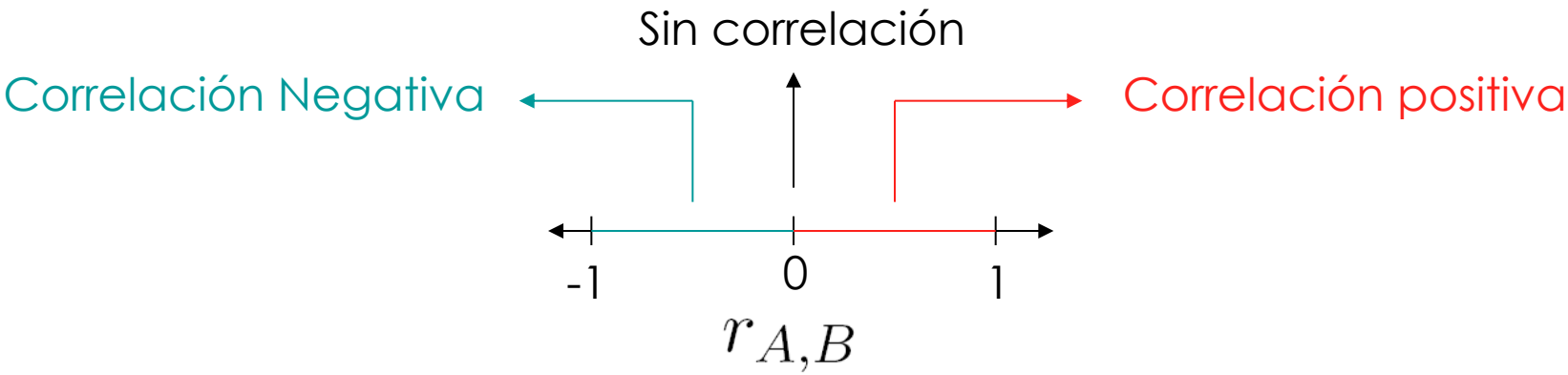
Redundancia

- Un atributo es redundante si puede ser derivado de otro. Errores en la identificación de la entidad suelen llevar a situaciones de redundancia
- Tablas no normalizadas también llevan a redundancias

- Puede ser detectada realizando un análisis de correlación:

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \overline{A})(b_i - \overline{B})}{N\sigma_A\sigma_B} \qquad -1 \leq r_{A,B} \leq 1$$

Redundancia (Cont.)



Detección y Resolución de conflictos entre valores

- Para la misma entidad, los valores del atributo proveniente de distintas fuentes de datos es diferente.

- Problema causado por diferencias de representación, escala, codificación, etc.

Detección y Resolución de conflictos entre valores (Cont..)

- **Diferencias en representación:** Por ejemplo, para una cadena de hoteles, el precio de una habitación en un hotel en distintas ciudades puede representar conceptos distintos, p.ej, uno incluye desayuno, impuestos u otro tipo de servicio que el precio en otra ciudad no contempla.

- **Diferencias en niveles de abstracción:** El total de ventas puede significar ventas totales en la cadena completa en una base de datos y en otra puede ser ventas totales en el hotel.

Detección y Resolución de conflictos entre valores (Cont..)

- **Diferencias de codificación:** Los valores para una entidad se nombran distintos en distintas fuentes de datos. Ej. Sexo ⟶ M,F ó 1,2 , etc.

- **Diferencias de escala:** Medidas en una base de datos pueden estar en centímetros, en otra en metros, etc.

- Al hacer matching entre atributos de diferentes fuentes de datos es necesario tener en cuenta la estructura de la información, ej. En una base de datos el descuento se aplicaba sobre un item y en otra el descuento se aplica sobre el total de la orden

Transformación

- Los datos se transforman y consolidan de tal forma de quedar listos para los procesos de minería de datos.
- La transformación puede envolver los siguientes procesos:

Smoothing: binning, regresión, clustering.

Normalización: Se modifica la escala de los atributos de tal forma que todos los valores queden dentro de un rango específico. Ej, 0 y 1, -1 y 1.

Construcción de características (feature construction): Nuevos atributos son construidos desde el mismo set de datos para mejorar el proceso de data mining.

Transformación(Cont..)

Agregación: Roll Up de algunos atributos (ventas mensuales, anuales, etc.)

Generalización: Datos son reemplazados por datos de niveles más altos (jerarquías de concepto). P. Ej, pasar de calles a comunas, ciudades a países, edad a un concepto más alto como “adulto-joven”, “senior”,etc.

Normalización

Útil en algoritmos que consideran distancias (KNN, clustering, etc.)

Normalización Min-Max: Realiza una transformación lineal en los datos originales. Si los rangos iniciales son *min_A y max_A,* y los rangos finales son *new_min_A* y *new_max_A,* la transformación de un valor v a un valor v' queda:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Normalización(Cont..)

Normalización z-score: Los valores para un atributo A son normalizados en base a la media y la desviación estándar:

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

Normalización(Cont..)

Normalización decimal (by decimal scaling)**:** Los valores para un atributo A son normalizados moviendo los puntos decimales:

$$v' = \frac{v}{10^j}$$

Donde *j* es el menor entero tal que $Max(|v'|) < 1$

Ej: rango inicial -971 a 990 ⟶ -0.971 a 0.990

Construcción de características

Se construyen nuevos atributos a partir de los existentes de tal forma de ayudar al proceso de data mining.
Ejemplo: Caso robo en cajeros automáticos

En algoritmos de clasificación se construyen inicialmente muchas características y luego en base a procesos de selección se dejan las mejores

Reducción de Datos

- A veces la cantidad de información hace que sea impracticable procesar la base de datos

- La idea es reducir la base de datos manteniendo la integridad en un alto porcentaje.

- Los algoritmos de minería de datos deben producir resultados muy similares en la base de datos reducida

Reducción de Datos

- Algunas estrategias de reducción son las siguientes:
  - Agregación del cubo de datos
  - Selección de un subconjunto de atributos
  - Reducción de dimensionalidad
  - Discretización (Bining, generación de rangos)
  - Jerarquías de concepto (Generalización).

Agregación del cubo de datos

- Algunos datos son agregados dependiendo de la información que se desea manejar.

- P. ej. Sólo se desean las ventas anuales, por lo tanto se deben sumar los datos mensuales.

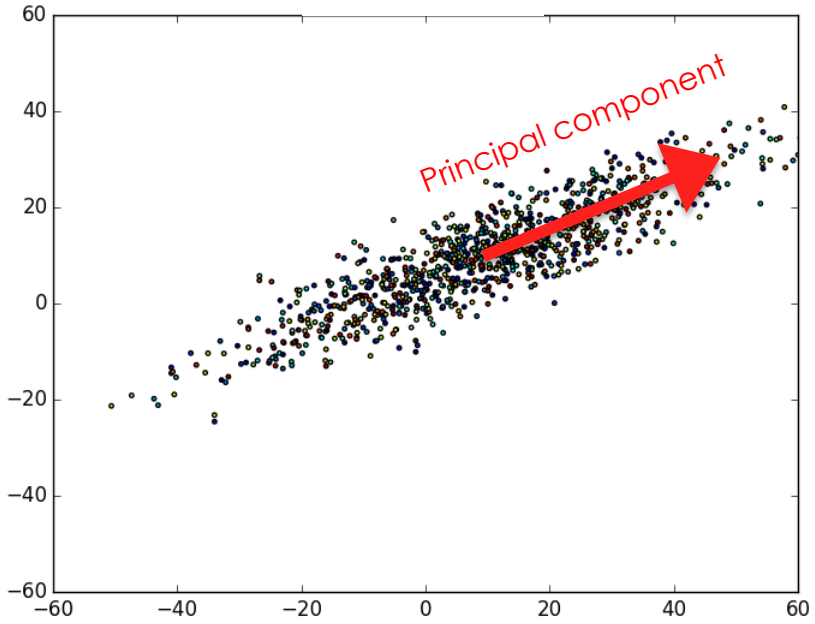
- Se logra una importante reducción de la cantidad de información sin pérdida de datos necesarios para el análisis

Reducción de Dimensionalidad

- Se aplican transformaciones de los datos de tal forma de obtener una versión comprimida de ellos.
- La versión comprimida tiene bajos porcentajes de pérdida de información
- Técnica comunmente utilizada: Principal Components Analysis (PCA).

Principal Components Analysis (PCA)

- Método de reducción de dimensionalidad
- Los atributos se transforman en un nuevo conjunto de atributos más reducido, donde cada uno de los atributos del conjunto nuevo es una combinación lineal de los atributos iniciales



Recordar algunos conceptos básicos necesarios

- Valor esperado de una variable x (E[x]) en palabras simples es el promedio de esa variable.
- Corresponde al promedio ponderado de ocurrencias de la variable.
- Los ponderadores corresponden a las probabilidades de ocurrencia de cada valor.
- Si todos los valores que puede tomar la variable son equiprobables, el valor esperado corresponde simplemente al promedio de las ocurrencias

Recordar algunos conceptos básicos necesarios

$$E[x] = \sum_i x_i \; p(x_i)$$ (Caso discreto)

$$E[x] = \int x \; p(x) dx$$ (Caso continuo)

- Si no conocemos p, pero tenemos **N** muestras de p:

$$E[x] \approx \frac{1}{N} \sum_i x_i$$

(Monte Carlo approximation)

$$x_i \sim p$$

muestran aparecen acorde con la distribución p

Conceptos básicos necesarios

- Valor esperado de una función **f** : (Importante en Machine Learning)

$$E[f(x)] = \int f(x)\, p(x) dx$$

- Si no conocemos p, pero tenemos **N** muestras de p:

$$E[f(x)] \approx \frac{1}{N} \sum_i f(x_i) \quad , \quad x_i \sim p$$

(Monte Carlo approximation)

Conceptos básicos necesarios

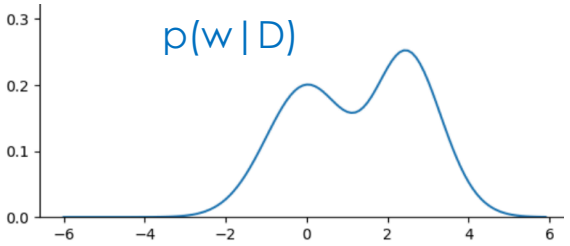
- Ejemplo: valor esperado de una función $\mathbf{f}$

Modelo predictor de la temperatura en una ciudad

Temp(d) = **w** * Temp (d-1) = $f(w)$

model data

Si w pasa a tener una distribución:



$$\text{Temp(d)} = \int f(w) * p(w|data)\, dw$$

Expected value of a function

Conceptos básicos necesarios

- Ejemplo: valor esperado de una función **f**

Si la distribución de w no se conoce, pero tenemos N casos:

$$\text{Temp(d)} = \frac{1}{N}\sum_i f(w_i)$$

Monte Carlo Approximation

$$w_i \sim p$$

Los casos vienen de la distribución p

Recordar algunos conceptos básicos necesarios

- Linealidad:

$$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$$

$$\mathbf{E}[aX] = a\,\mathbf{E}[X]$$

Recordar algunos conceptos básicos necesarios

- Varianza: Indicador que mide el grado de dispersión de los datos.
  - Corresponde al promedio de las diferencias al cuadrado con la media (el promedio de los datos)

promedio promedio

$$\operatorname{Var}(X) = \operatorname{E}\left[(X - \operatorname{E}[X])^2\right]$$

$$= \operatorname{E}\left[X^2 - 2X\operatorname{E}[X] + \operatorname{E}[X]^2\right]$$
$$= \operatorname{E}\left[X^2\right] - 2\operatorname{E}[X]\operatorname{E}[X] + \operatorname{E}[X]^2$$
$$= \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2$$

Recordar algunos conceptos básicos necesarios

- Covarianza:

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

$$= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$$

- Matriz de Covarianza:

$$\mu_i = \mathrm{E}(X_i)$$

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathrm{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathrm{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathrm{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}$$

Recordar algunos conceptos básicos necesarios

- Covarianza:

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y))$$

- Matriz de Covarianza (Notación matricial):

D x 1

D x D

$$\Sigma = \mathbf{E}\left[(\mathbf{X} - \mathbf{E}[\mathbf{X}])(\mathbf{X} - \mathbf{E}[\mathbf{X}])^{\mathrm{T}}\right]$$

$$\Sigma = \mathbf{E}(\mathbf{X}\mathbf{X}^{\mathrm{T}}) - \mathbf{E}(\mathbf{X})\,\mathbf{E}(\mathbf{X})^{\mathrm{T}}$$

Recordar algunos conceptos básicos necesarios

- Varianza: Algunas propiedades

$$Var(aX + b) = a^2 Var(X)$$

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$