DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Preprocesamiento de datos

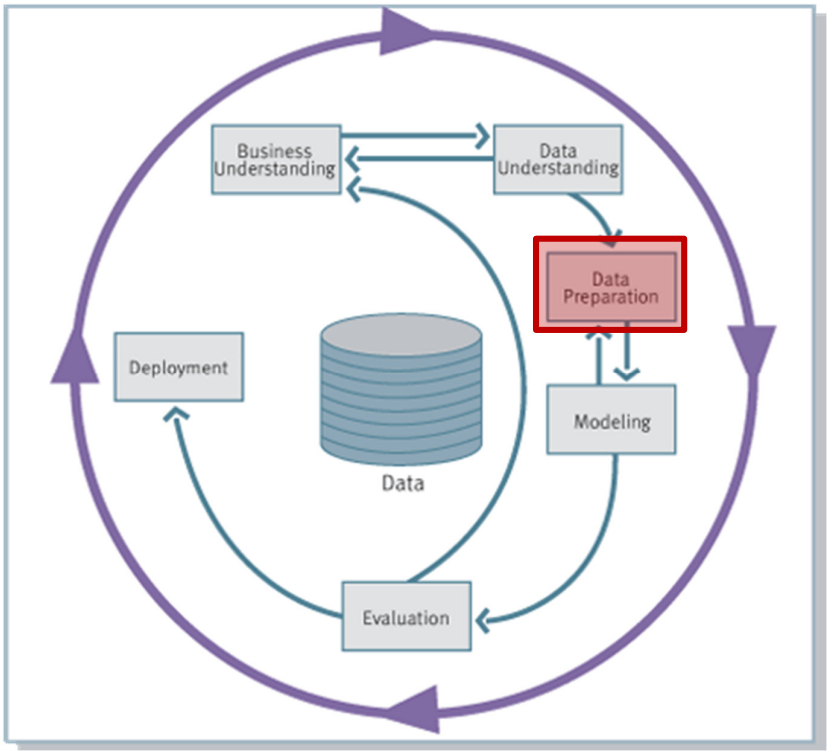
© Fernando Berzal, berzal@acm.org

Preprocesamiento de datos

- Datos
  - Fuentes de datos
  - Tipos de datos
- Exploración de datos
  [EDA: Exploratory Data Analysis]
  - Estadística Descriptiva
  - Visualización de datos
- Limpieza de datos
  - Valores nulos
  - Presencia de ruido
- Transformación de datos
  - Normalización y estandarización
  - Discretización
  - Extracción de características
- Reducción de datos
  - Técnicas de muestreo
  - Selección de características
- Ingeniería de características

Preprocesamiento de datos

Preprocesamiento de datos en CRISP-DM



Fuentes de datos

Fuentes de datos

- Bases de datos relacionales
- Bases de datos multidimensionales (DW)
- Bases de datos transaccionales
- Series temporales, secuencias y data streams
- Datos estructurados (grafos, redes sociales)
- Datos espaciales y espaciotemporales
- Textos e hipertextos (p.ej. Web)
- Bases de datos multimedia (p.ej. Imágenes)

Fuentes de datos

Bases de datos relacionales

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

no relation

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

Fuentes de datos

La estructura de datos más habitual para trabajar en minería de datos es un conjunto de datos en forma de tabla [dataset]:

- Cada fila de la tabla representa un caso, muestra, ejemplo, instancia, objeto o tupla.
- Cada columna de la tabla representa una variable, atributo, característica, dimensión o factor.
- Las variables pueden ser de muchos tipos.
- Puede haber datos "perdidos" (valores desconocidos y nulos).

Tipos de datos

Las variables o atributos de un conjunto de datos pueden ser…

Atributos discretos

- Conjunto finito o infinito contable de valores posibles.
- A menudo representados como valores enteros.
- Los atributos binarios son un caso particular.

Atributos continuos

- Números reales como valores.
- Medidos en la práctica con una precisión finita.
- Representados en coma flotante (32 o 64 bits).

Tipos de datos

Es importante ser conscientes de qué mide cada variable (medida en sentido estadístico) y de cómo se codifica el atributo (tipo de dato en sentido informático).

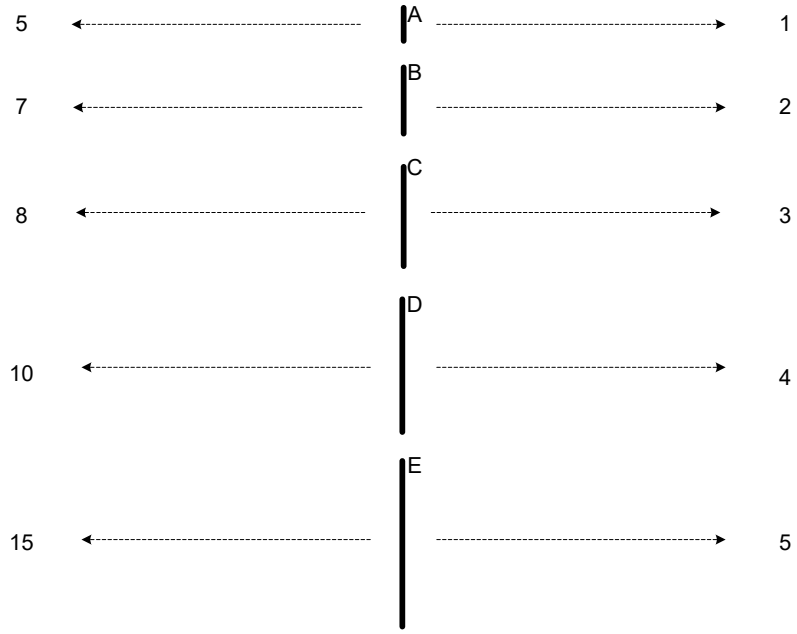
Tipos de medidas

- Categórica / cualitativa
  - Nominal (variable que toma valores no ordenados).
  - Ordinal (variable que toma valores ordenados).
- Numérica / cuantitativa / medida de escala
  - Intervalar (valores numéricos, resta con sentido).
  - Racional [ratio] (valores numéricos para los que tienen sentido operaciones de resta y división).

Tipos de datos

La forma en que se mide un atributo
puede no encajar bien con sus propiedades…

Esta escala preserva sólo el orden relativo de las longitudes medidas



Esta escala preserva la ordenación y la aditividad de las longitudes medidas

Tipos de datos

Datos categóricos / cualitativos

Atributos nominales

- ID (DNI)
- Código postal
- Color de ojos
- Sexo

Atributos ordinales

- Nivel de satisfacción (en una escala de 0 a N)
- Variable lingüística (p.ej. dureza baja, media, o alta)
- Calificación (suspenso, aprobado, notable, sobresaliente)
- Rankings de preferencias
- Números de calle en una dirección

Tipos de datos

Datos numéricos / cuantitativos

Atributos intervalares

- Fechas del calendario
- Temperaturas en grados Celsius (ºC) o Fahrenheit (ºF)

Atributos racionales

- Temperaturas en grados Kelvin (ºK)
- Frecuencias (número de ocurrencias de un evento)
- Tiempo transcurrido (p.ej. marcas deportivas, edad)
- Longitudes (pies, metros, millas, kilómetros o años-luz)
- Masas (gramos, onzas, libras, kilogramos o MeV/c^2)
- Cantidades de dinero (euros [€], dólares [$]…)

Tipos de datos

Atributo		Valores
Categórico / cualitativo	Nominal	Sólo se pueden distinguir (=,≠)
	Ordinal	También se pueden ordenar (<,>)
Numérico / cuantitativo	Intervalar	Las diferencias tienen sentido (+,-)
	Racional	Diferencias y ratios con sentido (+,-,*,/)

Tipos de datos

Atributo		Operaciones
Categórico / cualitativo	Nominal	Moda, entropía, contingencia, correlación, test χ^2
	Ordinal	Mediana, percentiles, correlación de rangos, tests de signo
Numérico / cuantitativo	Intervalar	Media aritmética, desviación estándar, correlación de Pearson, tests t y F
	Racional	Media geométrica, media armónica, variación porcentual

Tipos de datos

Atributo		Transformaciones válidas
Categórico / cualitativo	Nominal	Permutaciones
	Ordinal	Cambio de valores que preserve el orden, i.e. f(x) con f monótona
Numérico / cuantitativo	Intervalar	Transformaciones lineales f(x) = a*x + b
	Racional	Cambios de escala (unidades) f(x) = a*x

Tipos de datos

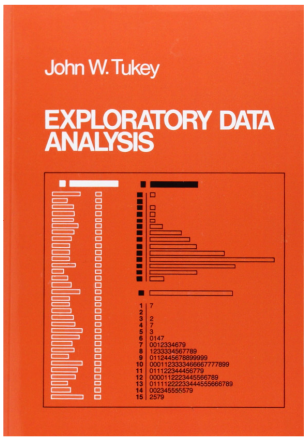
Nivel de medida	Tipo de datos			
	Numérico	Cadena	Fecha	Tiempo
Escala		n/a		
Ordinal				
Nominal				




Exploración de datos

Análisis exploratorio de datos
[EDA: Exploratory Data Analysis]

John Tukey: "Exploratory Data Analysis," Addison-Wesley, 1977. ISBN 0201076160.



- Permite formular hipótesis iniciales.
- Ayuda a elegir qué herramientas y técnicas resultan más adecuadas para preprocesar y analizar un conjunto de datos.
- Enfocado a la visualización de datos (explota la capacidad del ser humano para extraer patrones).

Estadística descriptiva

OBJETIVO

Descripción básica de los datos
para facilitar su comprensión

Estadísticos

Medidas de resumen, que se calculan a partir de los propios datos y se usan para obtener una visión global de la distribución de los datos.

- Tamaño muestral (N), número de casos en la muestra.
- Estadísticos de localización (dan una idea de cuáles son los valores habituales de la distribución).
- Estadísticos de dispersión (dan una idea de cuál es la variabilidad de los datos).

Estadística descriptiva

Distribución de frecuencias

- Frecuencia absoluta: n(x)
  Número de veces que aparece un valor x.
- Frecuencia relativa: f(x) = n(X) / N
  (se puede dar también en porcentajes, 100*f(x))
- Frecuencia acumulativa: $F(x) = \Sigma_{y \leq x} f(y)$
- Frecuencia acumulativa absoluta: $Fa(x) = \Sigma_{y \leq x} n(y)$

Estadística descriptiva

Distribución de frecuencias

- **Percentiles**
  p_s = max { x | 100*f(x) ≤ s } con s de 0 a 100

- **Cuartiles**
  - Mínimo (percentil 0)
  - Primer cuartil Q_1 (percentil 25)
  - Segundo cuartil (percentil 50), a.k.a. mediana
  - Tercer cuartil Q_3 (percentil 75)
  - Máximo (percentil 100)

Estadística descriptiva

Medidas de tendencia central

Para datos numéricos…

- **Media aritmética** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\mu = \frac{\sum x}{N}$

- **Media ponderada** $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

- **Media truncada** [trimmed mean],
  sin valores extremos, p.ej. sin el 5%, de $p_{2.5}$ a $p_{97.5}$

Estadística descriptiva

Medidas de tendencia central

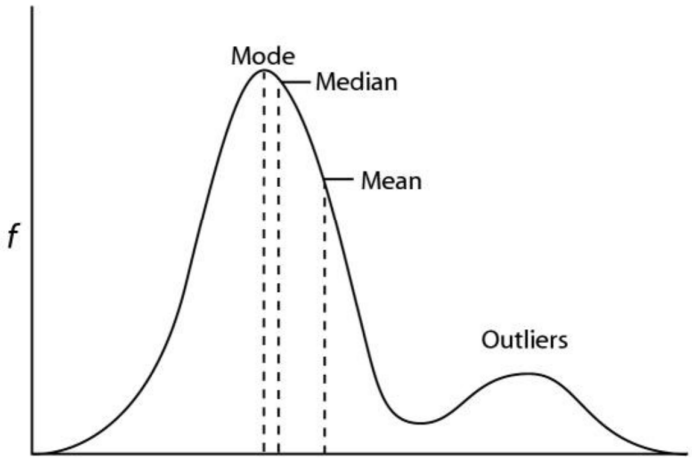
Para todo tipo de datos…

- **Mediana** (percentil 50)
  - Cálculo: Se ordenan todos los valores y se escoge el central. Si el número de valores es par, se toma la media aritmética de los dos valores centrales.
  - Menos sensible a outliers que la media aritmética.

- **Moda** (valor más frecuente)
  - Si la distribución tiene una única moda, se dice que la distribución es unimodal (si no, multimodal).

Estadística descriptiva

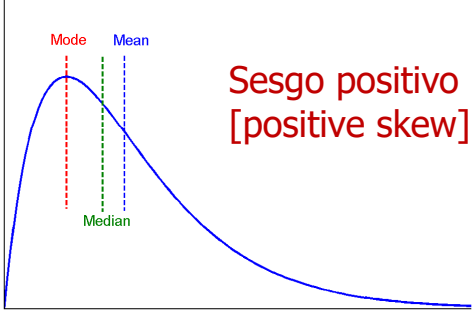
Medidas de tendencia central
(estadísticos de localización)

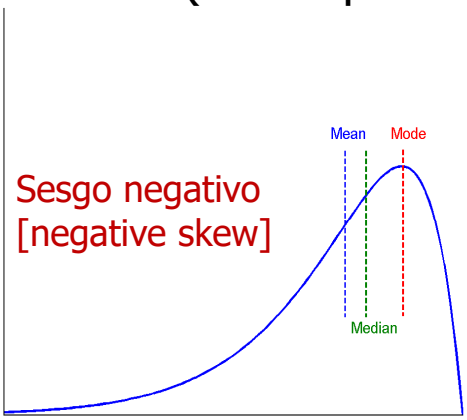


Estadística descriptiva

Sesgo de la distribución de los datos

- Distribución simétrica:
  media = mediana = moda (si es unimodal)
- Distribución sesgada positivamente (a la derecha)
  media ≥ mediana ≥ moda
- Distribución sesgada negativamente (a la izquierda)
  media ≤ mediana ≤ moda





Estadística descriptiva

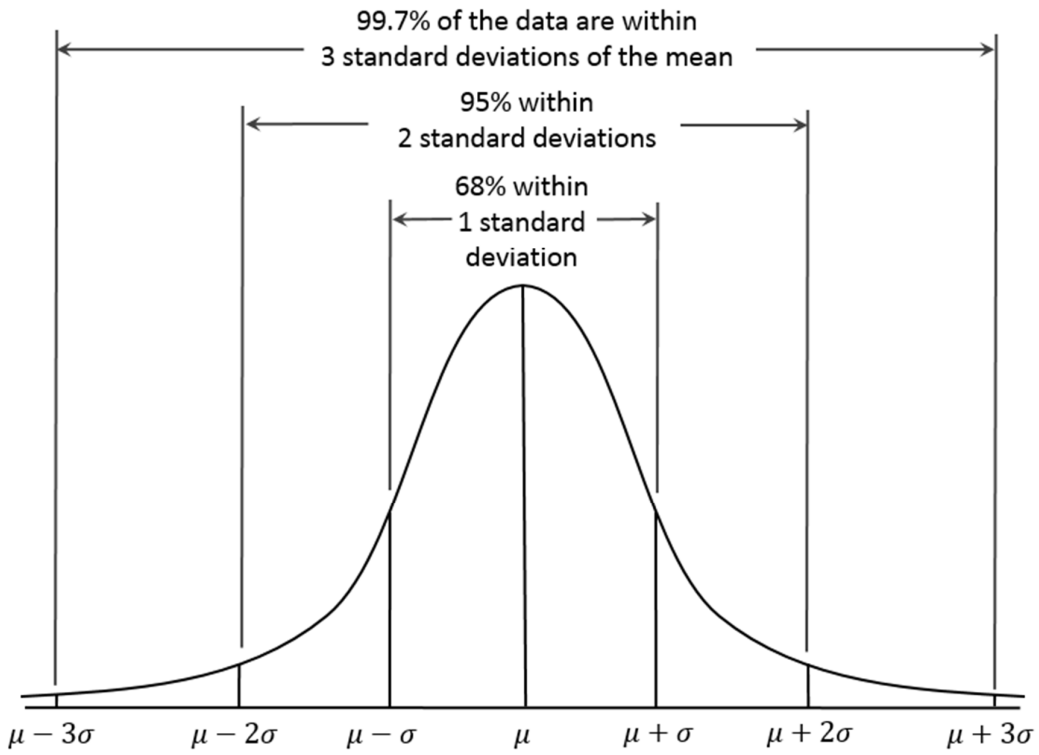
Medidas de dispersión

- Varianza $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i{}^2 - \mu^2$
- Desviación típica (raíz cuadrada de la varianza) σ
- Media de la desviación absoluta $AAD = \frac{1}{N}\sum_{i=1}^{N}|x_i - \mu|$
- Mediana de la desviación absoluta $MAD = p_{50}\{|x_i - \mu|\}$
- Rango intercuartil $IQR = p_{75} - p_{25}$

Estadística descriptiva

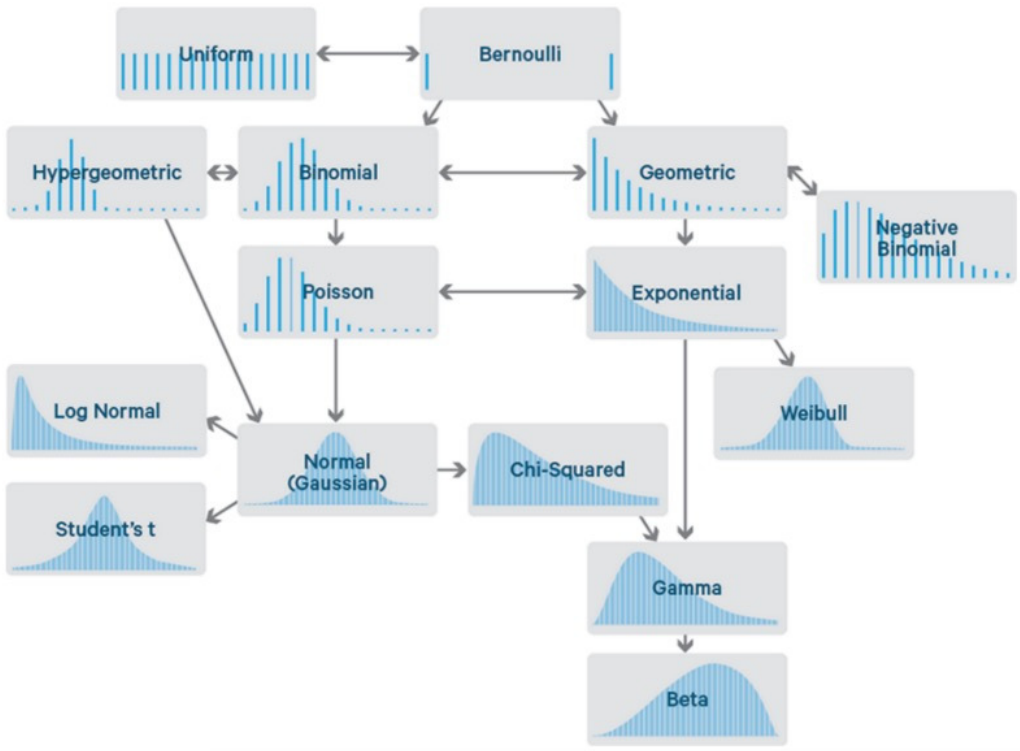
La distribución normal

$$N(\mu,\sigma) \quad f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu-x)^2}{2\sigma^2}}$$



Estadística descriptiva

Otras distribuciones de probabilidad



Estadística descriptiva

Divergencia KL [Kullback-Leibler]
para comparar dos distribuciones de probabilidad.

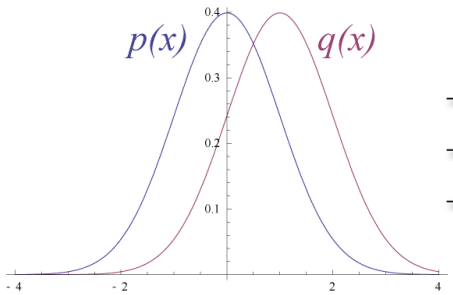
$D_{KL}(p(x) \,||\, q(x))$

- Basada en teoría de la información:
  Información perdida cuando se utiliza q(x) para aproximar p(x)

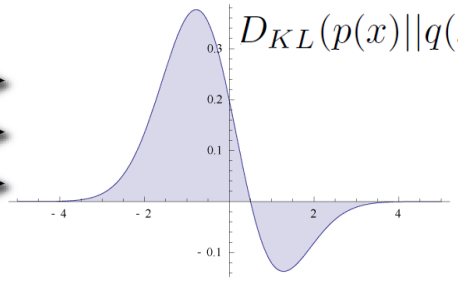
$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

Estadística descriptiva

Divergencia KL [Kullback-Leibler]
para comparar dos distribuciones de probabilidad.

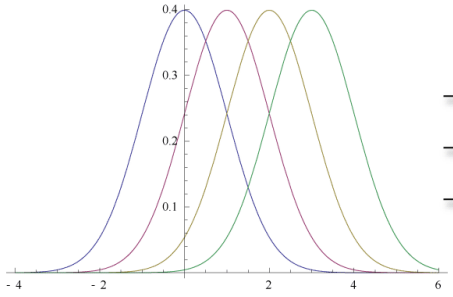


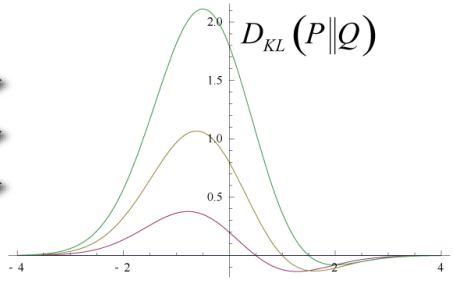
Original Gaussian PDF's



KL Area to be Integrated

$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$





Estadística descriptiva

Medidas de correlación (para atributos categóricos)

Test χ^2 [chi-squared]

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

observed

expected

- Hipótesis nula:
  Las dos distribuciones son independientes.
- Los ejemplos que más contribuyen al valor de χ^2 son aquéllos cuya frecuencia es más diferente de la esperada.
- Cuanto mayor sea el valor de χ^2, más probable es que las variables estén relacionadas.

Estadística descriptiva

Medidas de correlación (para atributos categóricos)

Test χ^2

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(column)	300	1200	1500

- Los números entre paréntesis son los valores esperados.

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- Jugar al ajedrez y que le guste la ciencia ficción son valores correlados en nuestro conjunto de datos:
  se rechaza la hipótesis nula al nivel de confianza 0.001

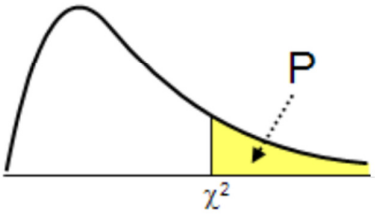
Estadística descriptiva

Medidas de correlación (para atributos categóricos)

Test χ^2

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

Values of the Chi-squared distribution



	P										
DF	0.995	0.975	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458

- **Grados de libertad [DF: degrees of freedom]**
  Número de valores que pueden variar libremente.
  DF = (#categorías$_{Ajedrez}$ – 1) * (#categorías$_{SF}$ – 1).

Estadística descriptiva

Medidas de correlación (para atributos numéricos)

- **Covarianza**
  Generalización de la varianza

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

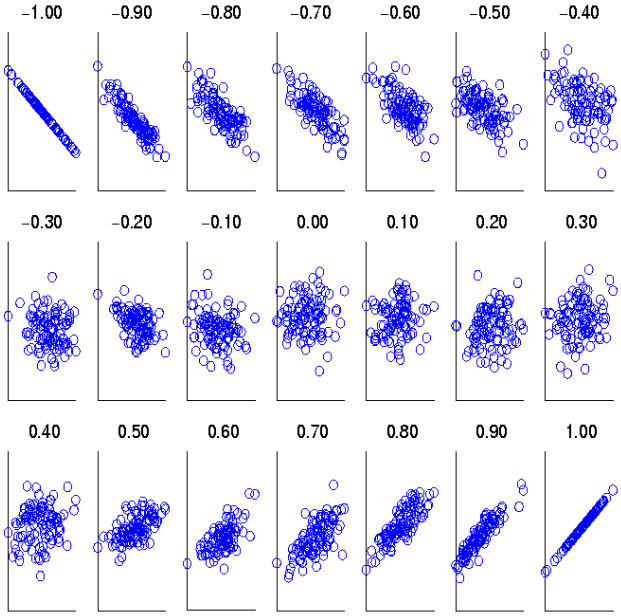
- **Correlación**
  (normalizada entre -1 y 1)

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Si las variables son independientes,
covarianza y correlación son nulas (a la inversa, no)

Estadística descriptiva

Medidas de correlación (para atributos numéricos)



Correlación negativa

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Correlación positiva

Variando el coeficiente de correlación de -1 a +1

Estadística descriptiva

Matriz de covarianza

- Las varianzas y covarianzas de dos variables se pueden representar en una matriz 2x2:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

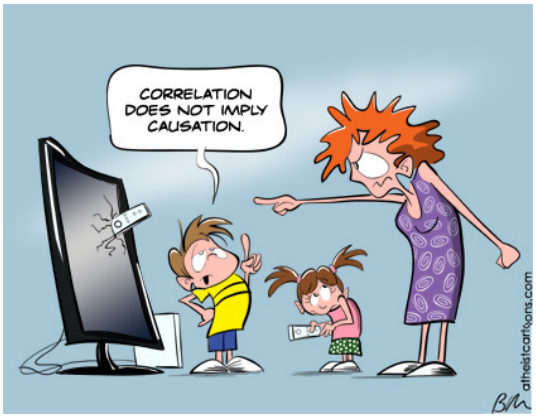
- Generalizando para d dimensiones/variables, obtenemos la matriz de covarianza:

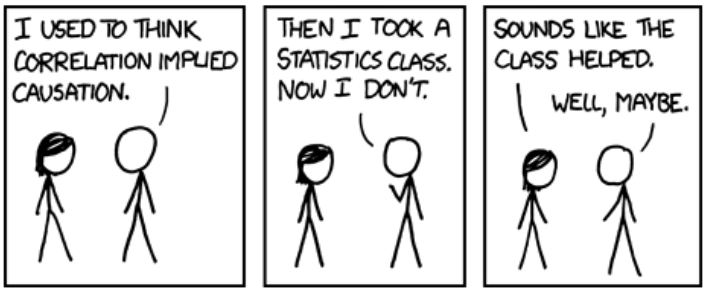
$$\boldsymbol{\Sigma} = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

Estadística descriptiva

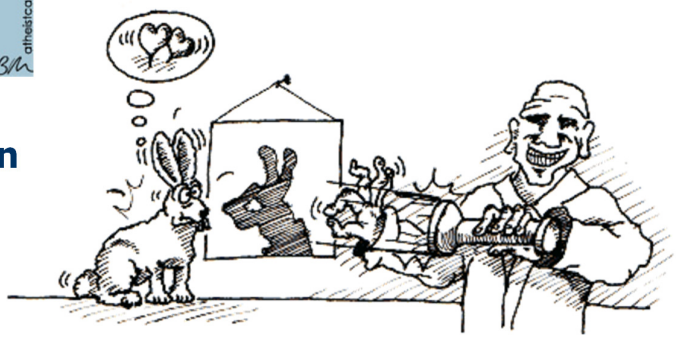
Medidas de correlación

Correlación no implica causalidad





"Correlation is not causation but it sure is a hint."
-- Edward Tufte

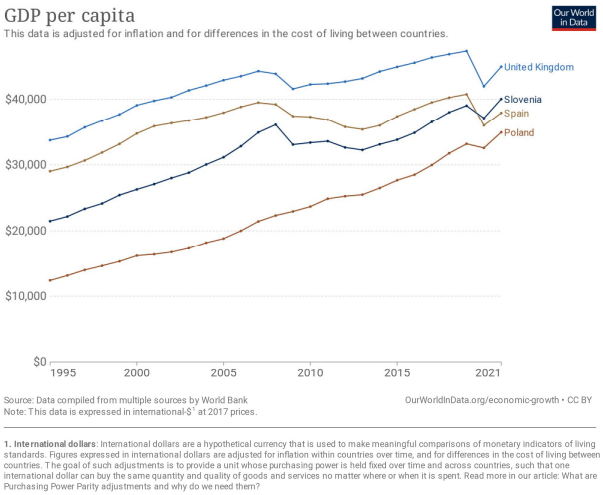


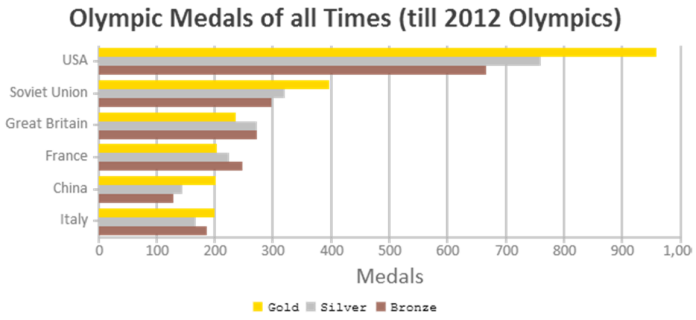
Visualización de datos

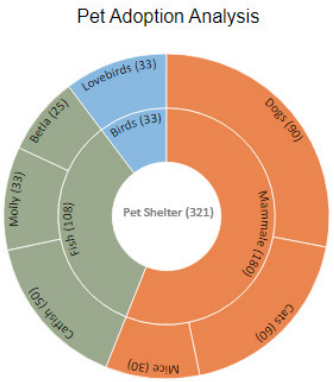
- Diagramas de líneas, barras y sectores
- Histograma de frecuencias (absolutas, relativas, acumuladas)
- Diagrama de cuartiles & Q-Q
- Diagrama de cajas [boxplot]
- Nube de puntos [scatter plot]

Visualización de datos

Diagramas de líneas, barras y sectores
[line, bar & pie charts]

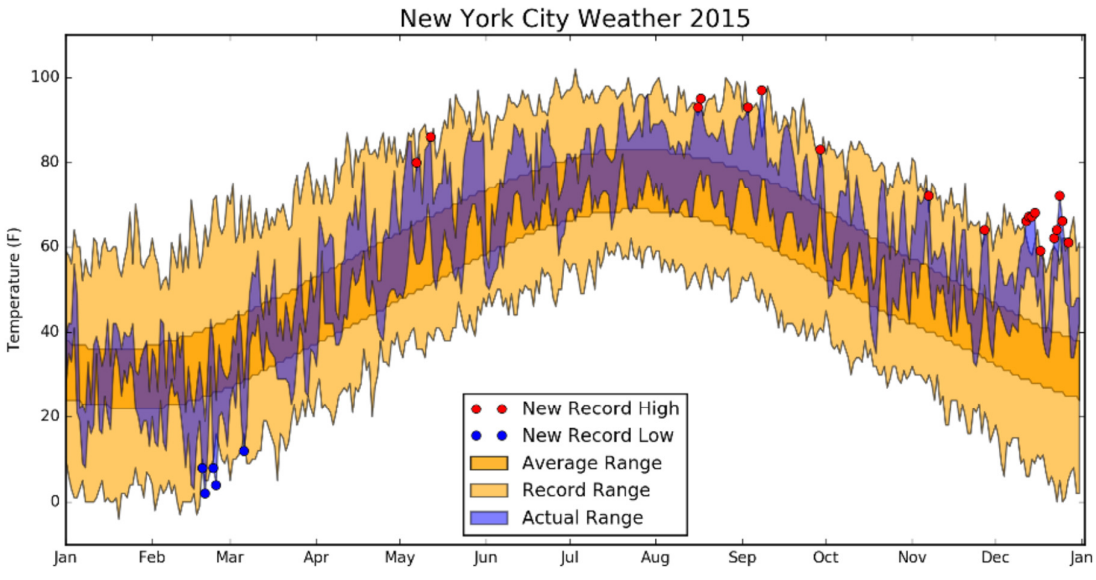


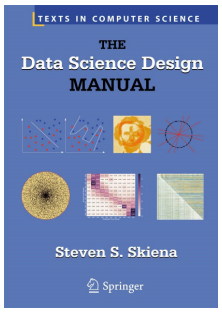




Visualización de datos

Ejemplo: Temperaturas anuales en Nueva York

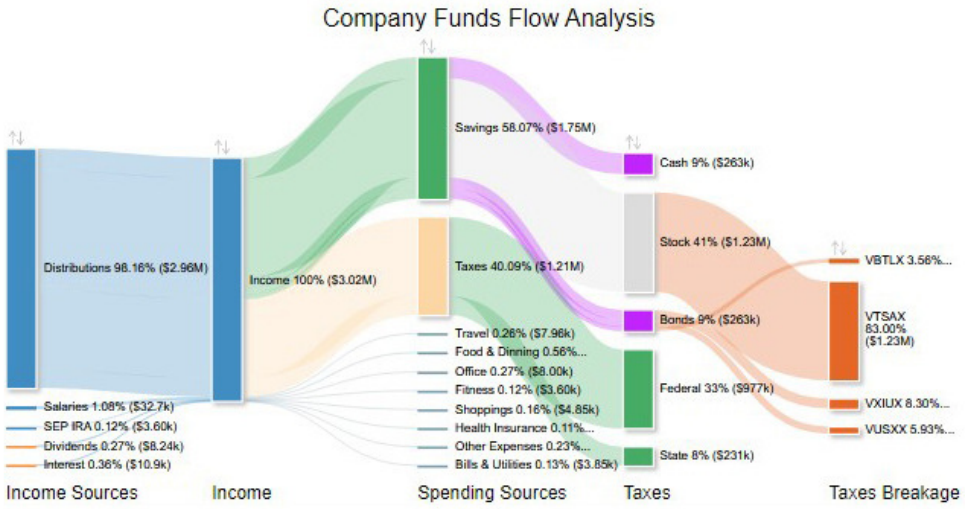




Visualización de datos

… y otros no tan estándar:

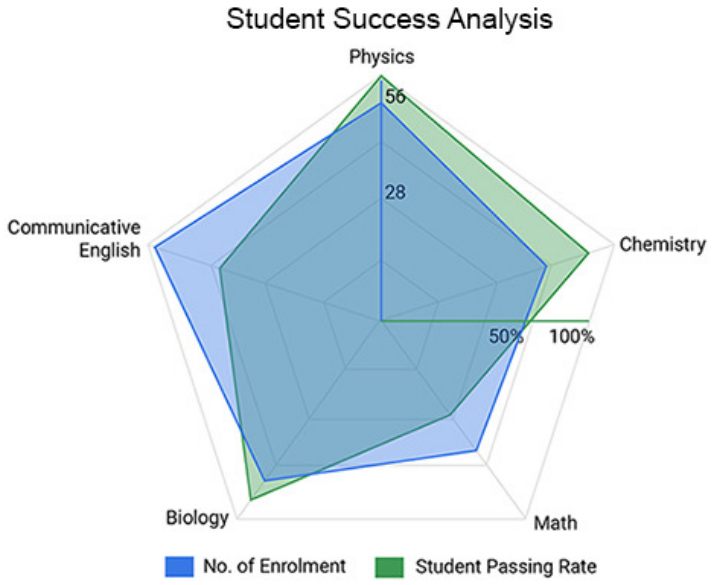
Sankey chart (flujos)



Visualización de datos

… y otros no tan estándar:

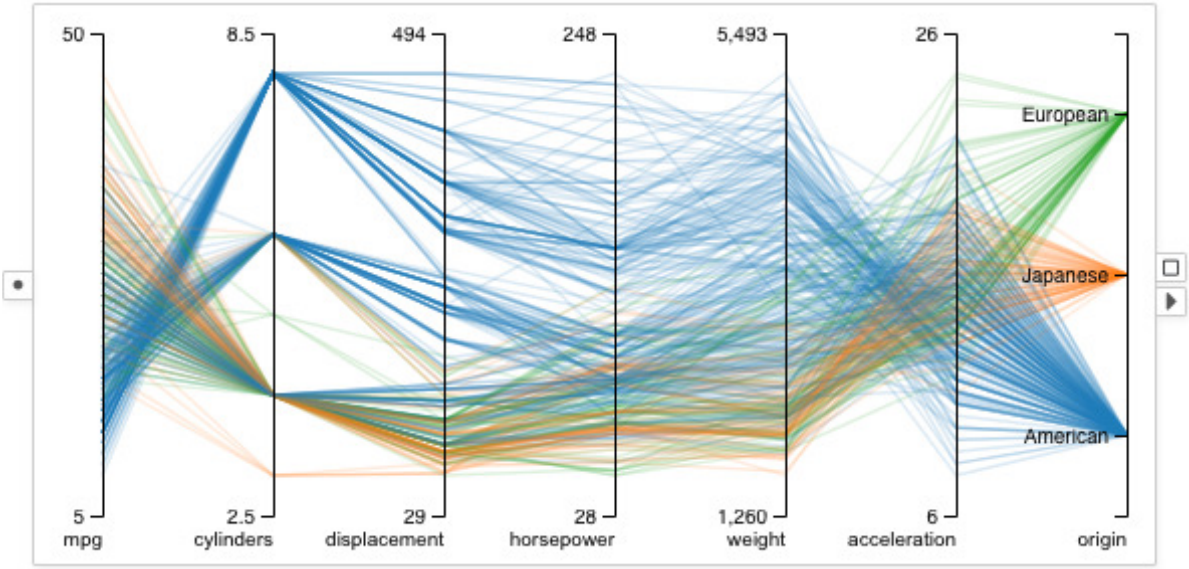
Diagrama de Kiviat (a.k.a. radar|cobweb|spider chart)



Visualización de datos

… y otros no tan estándar:

Coordenadas paralelas (Alfred Inselberg)

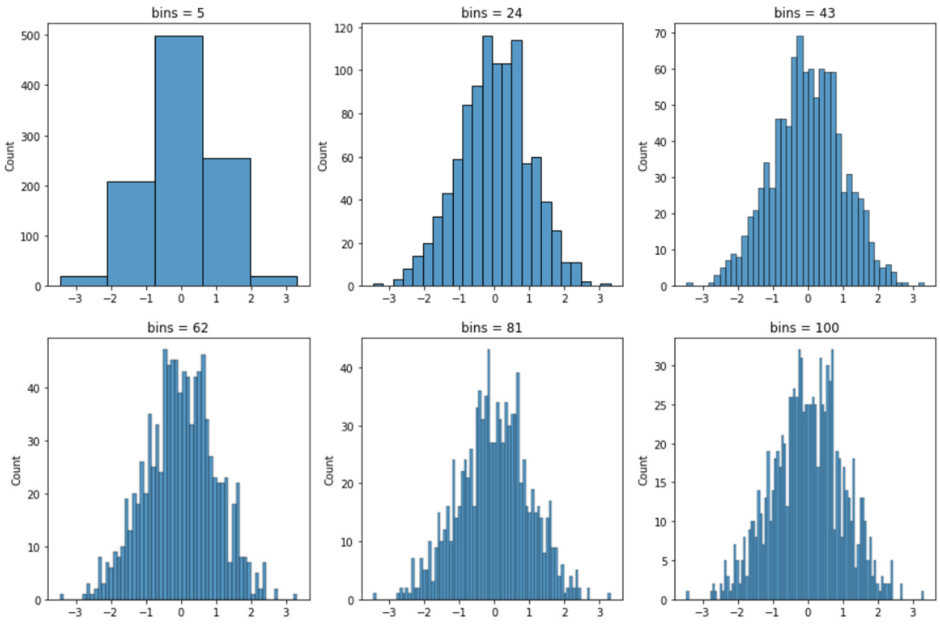


Alfred Inselberg: "The Plane with Parallel Coordinates". *Visual Computer* **1**(4):69–91, 1985. DOI 10.1007/BF0189835

Visualización de datos

Histograma

de frecuencias absolutas, relativas o acumuladas

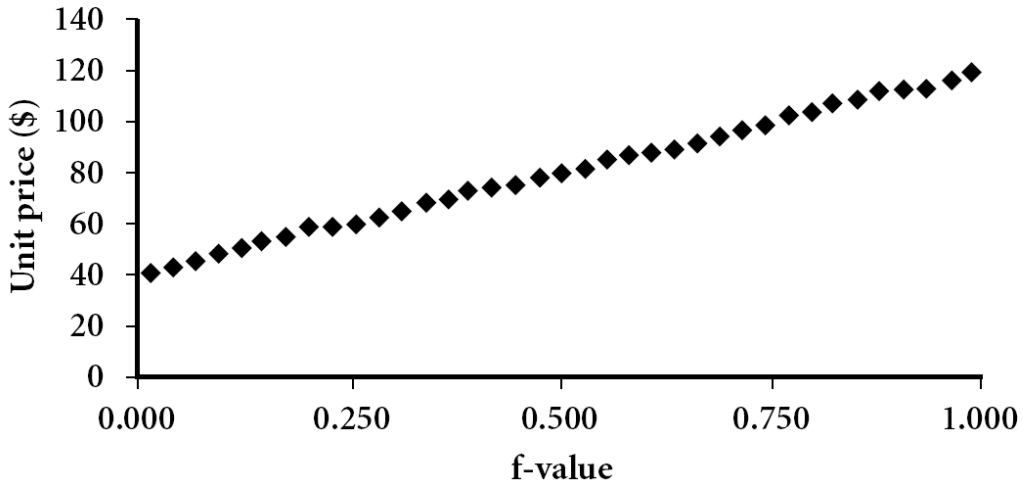


Con diferentes números de intervalos

Visualización de datos

Diagrama de cuantiles [quantile plot]

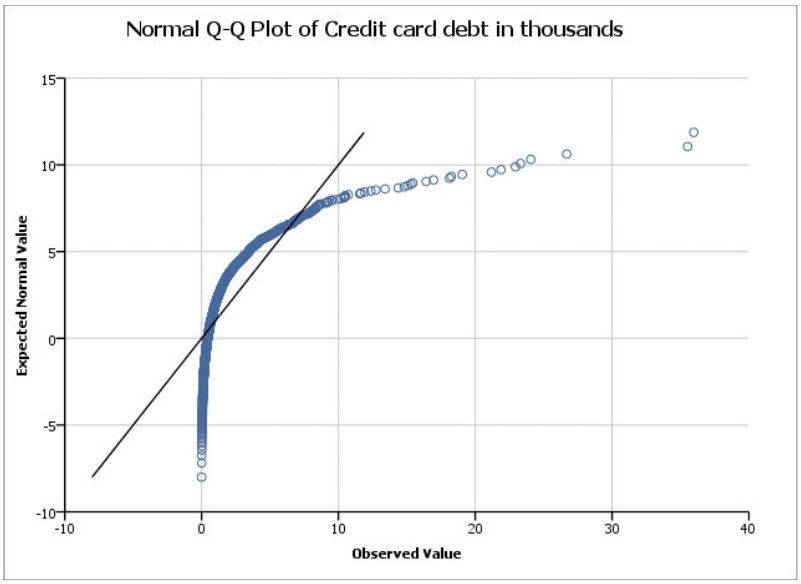
como un histograma de frecuencias acumuladas



Visualización de datos

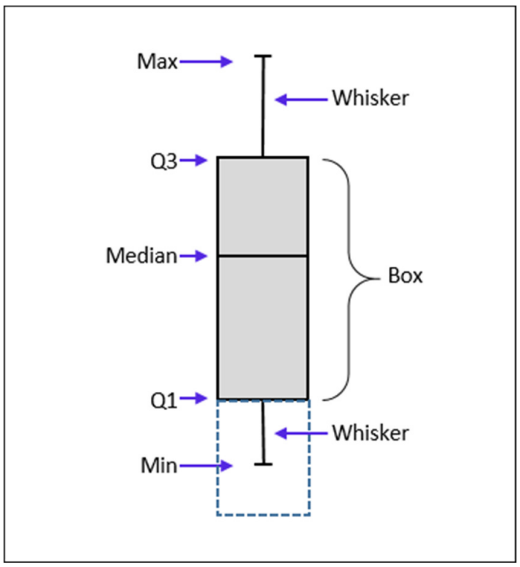
Diagrama Q-Q [quantile-quantile plot]

muestran los cuantiles de una distribución frente a los de otra (para observar diferencias)



Visualización de datos

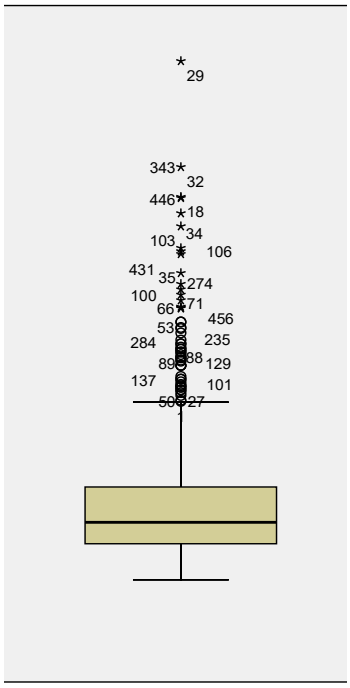
Diagrama de cajas [box plot]
Resumen de la distribución de los datos, propuesto por Tukey

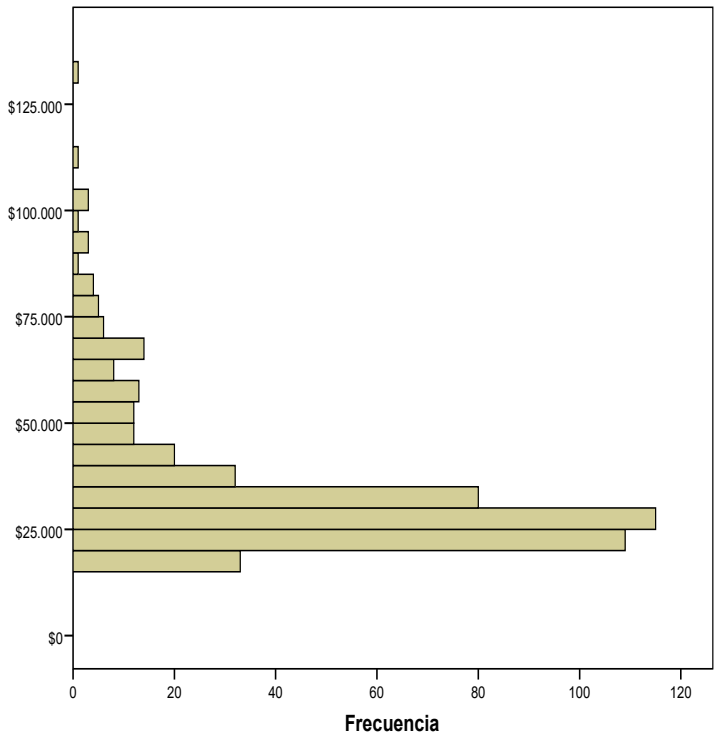


- Muestra los cuartiles y el rango intercuartil (IQR).
- Las "patillas" de la caja pueden representar mínimo y máximo, percentiles (p.ej. 10-90) o un múltiplo del IQR (p.ej. 1.5) para mostrar de forma explícita la existencia de **outliers**.

Visualización de datos

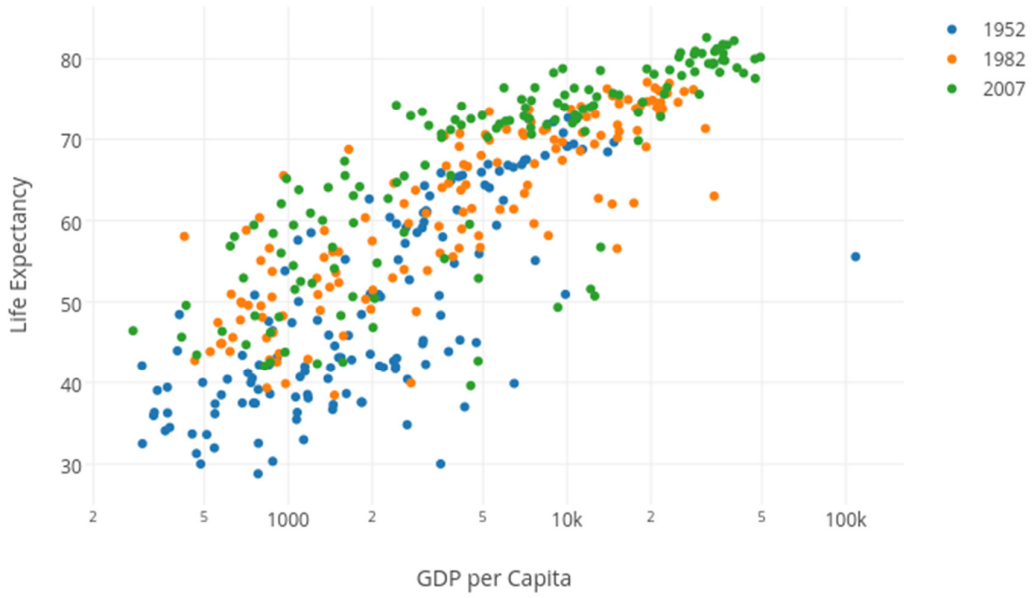
Diagrama de cajas [box plot]





Visualización de datos

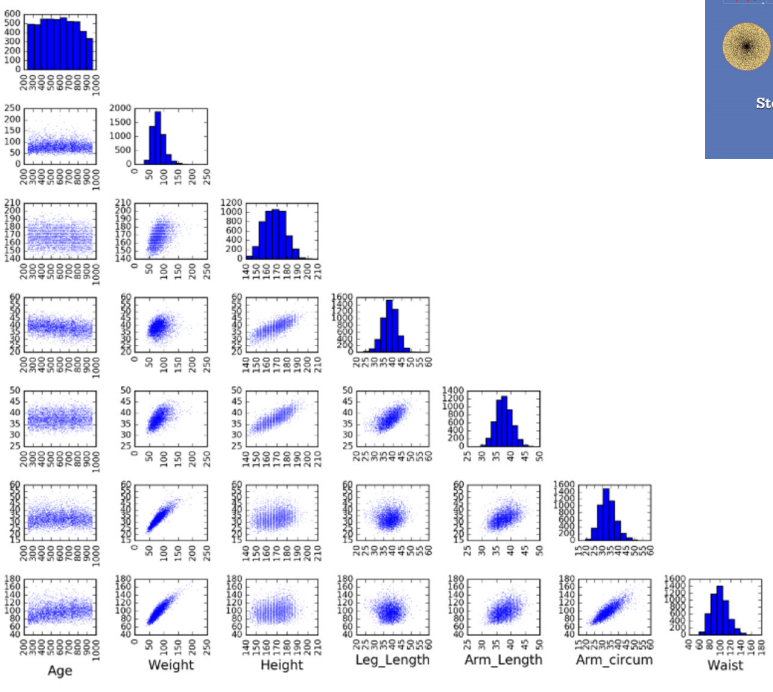
Nube de puntos [scatter plot]

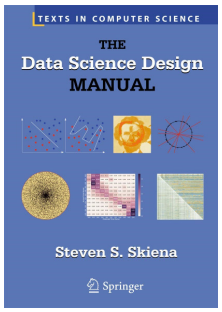


Visualización de datos

Nubes de puntos [scatter plots]

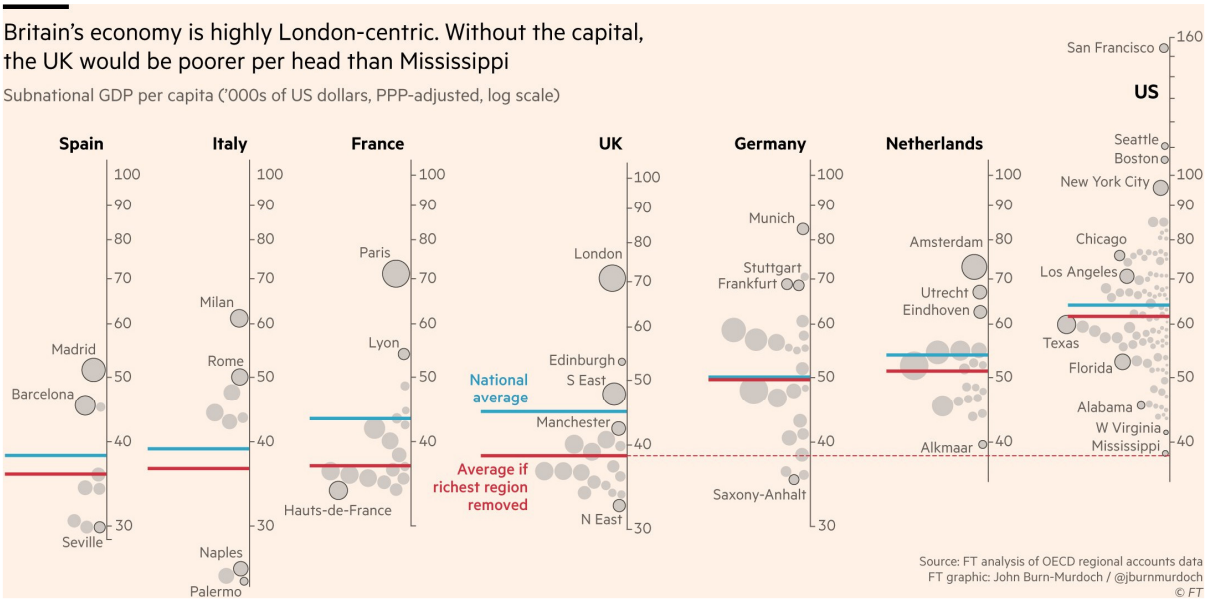
Pares de variables (vista rápida de distribuciones y correlaciones)





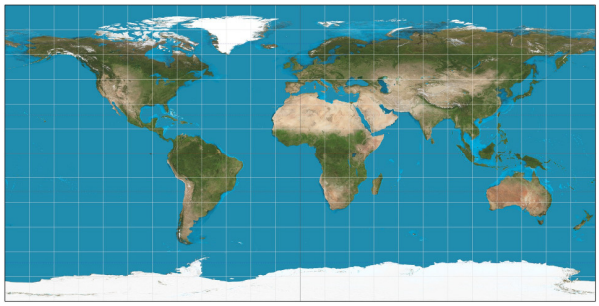
Visualización de datos

Combinación de múltiples técnicas…



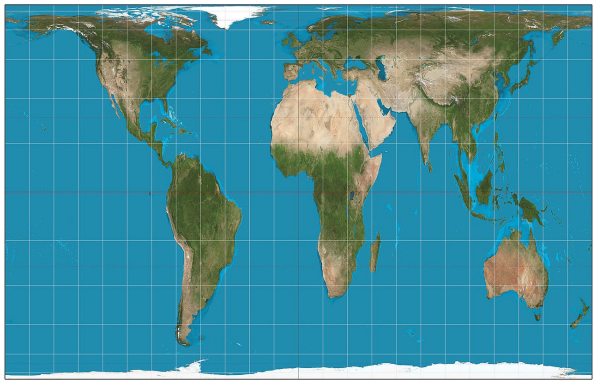
Visualización

¡OJO! El uso de áreas puede ser problemático…



Mercator

(líneas de rumbo rectas,
para ayudar en la navegación)



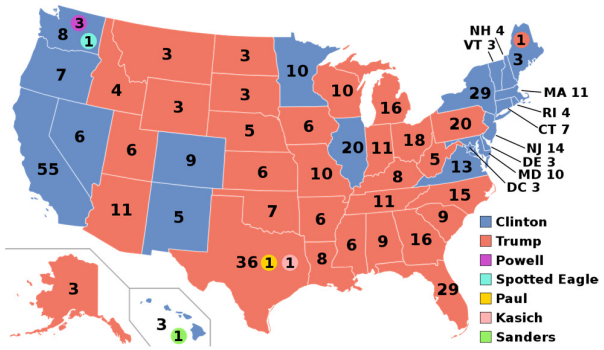
Gall-Peters

(respeta las áreas relativas)

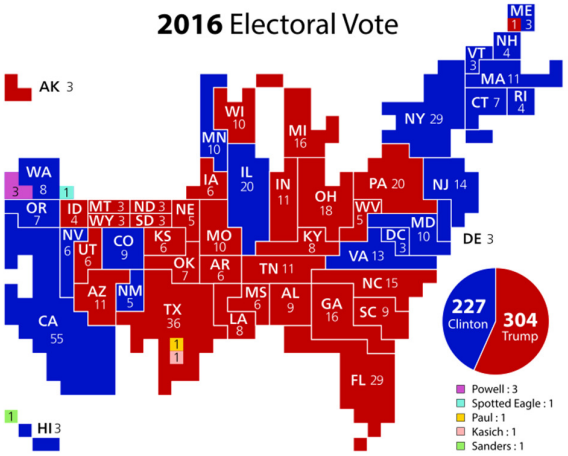
Dos proyecciones cilíndricas

Visualización

… pero también se puede aprovechar

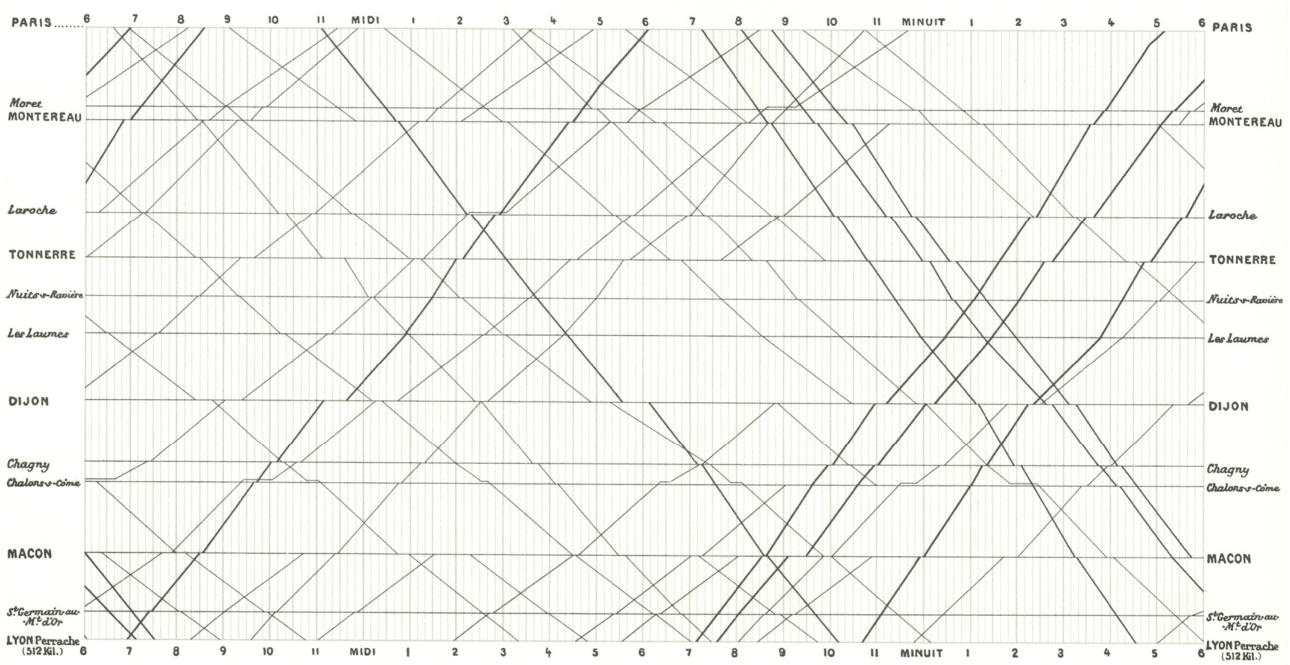


Mapa



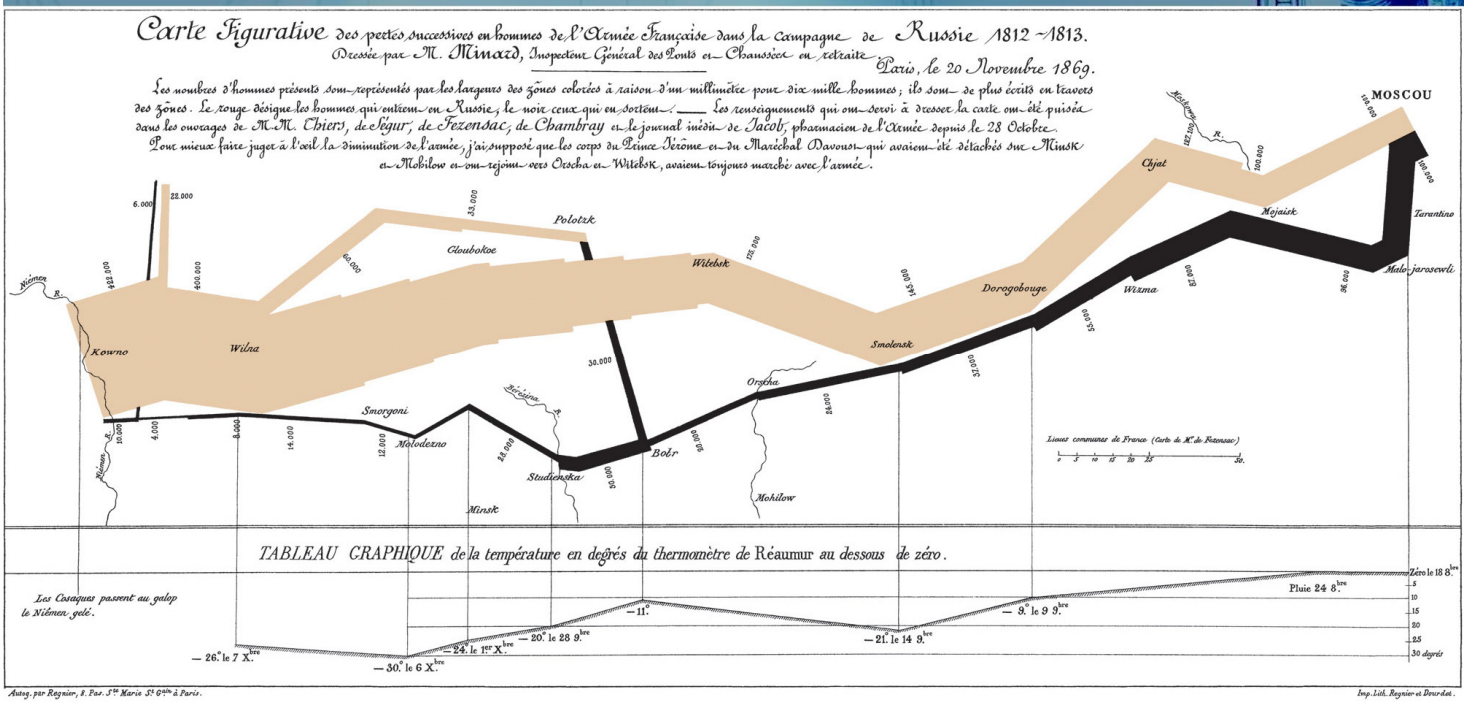
Cartograma

Visualización de datos



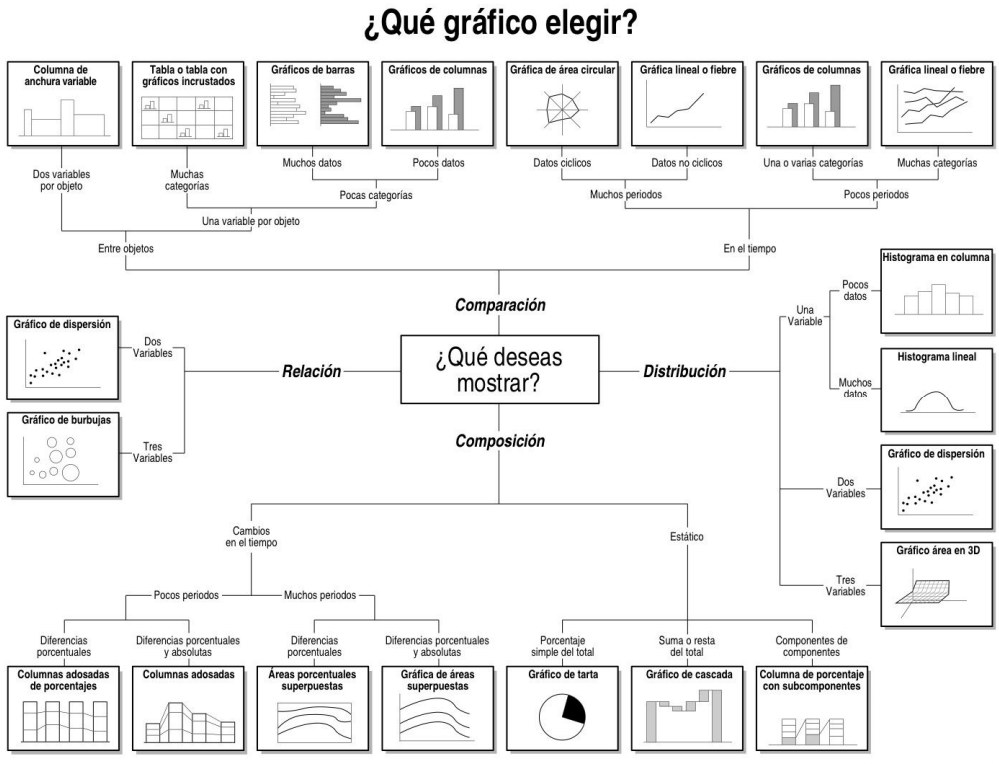
Un ejemplo famoso:
Horario de trenes de Marey (1885)

Visualización de datos



Un ejemplo famoso: el mapa de Minard de 1869 de la campaña de Napoleón en Rusia (1812-1813)

Visualización de datos

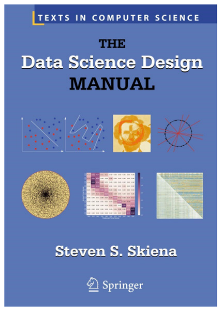


Traducción a cargo de Víctor Caballero con autorización del autor
www.Vectart.com / contact@vectart.com

www.ExtremePresentation.com
© 2009 A. Abela - a.v.abela@gmail.com

Visualización de datos

Consejos de Skiena



- Maximizar el uso de tinta para mostrar datos (tinta para datos / tinta total del gráfico)
- Minimizar el factor “mentira” [lie factor], i.e. tamaño de un efecto en el gráfico / tamaño del efecto en los datos.
- Minimizar efectos visuales que nada tienen que ver con los datos [chartjunk].
- Usar escalas adecuadas y etiquetas claras.
- Realizar un uso efectivo del color.
- Aprovechar el poder de la repetición para facilitar la realización de comparaciones.

Limpieza de datos

La calidad de los datos afecta a los resultados que se pueden obtener usando técnicas de minería de datos.

1957

Work With New Electronic 'Brains' Opens Field for Army Math Experts

WASHINGTON — BIZMAC UNIVAC, GARBAGE IN-GARBAGE OUT—all new terms in the Army.

They haven't replaced GOVERNMENT ISSUE or BY THE NUMBERS and probably won't, but in a small, select circle they're heard more frequently than the old expressions used by millions of veterans of World War II and Korea.

For these colorful expressions are part of the work-day vocabulary of the military mathematicians who man the Army's electronic computors, the brains that fed the right figures by the men who know how come up with the right answers in a matter of split seconds.

Specialist 3 C William D. Mellin of Gardner, Mass., is a master of one of these machines, some of which are as big as four basket ball courts.

• • •

MELLIN, a man of unusual occupation, has an unusual title. He's an electronic computer programmer, one of a group performing as highly specialized soldiers in this unique military field.

In his Signal Corps work with the computer Mellin might be given a problem to solve in connection with radio frequency data. Done the old fashioned way it might have taken 50 girls with hand calculating machines a week to come up with the answer.

One man however, in this case Mellin, "programs the problem, a pre-planning task that may take a few hours or a couple of days of figuring out the step by step process to be taken by the electronic brain."

When this is completed the information is fed, often as mathematical formulae, into the maw of the machine and depending on the magnitude of the problem the answer is obtained in seconds or minutes.

• • •

NEXT STEP is debugging or checking the accuracy of the work. Did the results meet the original requirements or should the problem be rerun with corrections?" That's where "Garbage in Garbage-out" (GIGO) enters the vocabulary.

Mellin explains: "If the problem has been sloppily programmed the answer will be just as incorrect. If the programmer made mistakes the machine will make mistakes. It can't correct them because it can't do one thing. It can't think for itself."

Men like Mellin are not picked out of basic training units and taught their Army jobs by the numbers. But the Army taught him plenty just the same. A 1952 Duke University graduate in economics and mechanical engineering, he became interested in computers as he watched the television reports of that year's national elections.

Fascinated by the possibilities offered in the relatively new field he went to work for one of the 'brain' manufacturers, his starting job consisting of eight hours of school daily followed by six months apprenticeship on the machine.

• • •

WHEN MELLIN entered the Army he was quickly returned to several schools operated by manufacturers of various models that the Army uses. The idea, of course, was to help him become familiar with the different machines used by the Army and especially by the Signal Corps.

All the 'brains' incidentally aren't found in the Pentagon or in the Army's research and development laboratories. They've also become almost indispensable in what old soldiers still fondly call "the field." The employment of new weapons systems requires the use of lightning calculations which has created a demand for trained programmers.

The complexity of the modern Army means equally complex supply and accounting systems to insure that field units have what they need when and where they need it. This means a smooth flow of weapons and equipment from manufacturer to combat soldier. Today that flow is made possible without miles of files.

Today one programmer "brain"-team can keep track of more than 100 million facts and then file all this information in a couple of reels of magnetic tape.

A 'BRAIN'S' BRAIN — Mechanical brains are useless without first-class human mind to figure what information to feed them. Specialist Third Class William D. Mellin of Gardner, Mass., does this work for the Army, which uses several types of electronic computers in supporting headquarters and combat unit activities.

Observe Anniversary

SALT LAKE CITY, Utah (AP) — The famed Salt Lake Tabernacle choir is observing an anniversary — its 25th successive year of Sunday morning broadcasts over the Columbia Broadcasting System. President Eisenhower and a score of other government officials have sent congratulations.

Limpieza de datos

Problemas con la calidad de los datos

- Datos incorrectos e imprecisos
  (presencia de errores e imprecisiones)
- Datos perdidos
  (falta de datos).
- Datos duplicados
  (provenientes de la integración de datos)
- Datos desequilibrados
  (muchos de un tipo, muy pocos de otro)
- Anomalías [a.k.a. outliers]
  (distorsionan los resultados del análisis de datos)
- Desfase temporal [a.k.a. timeliness]
  (datos no actualizados)

Limpieza de datos

Los datos disponibles…

- … pueden ser sólo aproximados.
- … pueden incluir errores (o ser fake ;-).
- … pueden incluir ruido y anomalías [outliers].
- … pueden ser insuficientes (valores perdidos).
- … pueden venir duplicados.

Limpieza de datos

Datos incorrectos e imprecisos

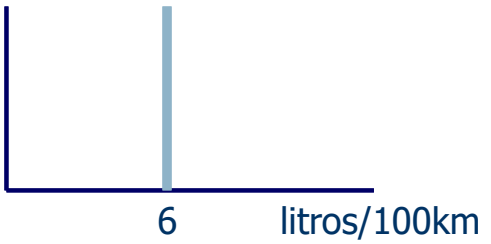
Diferencias entre el valor del dato
y el verdadero valor del atributo medido

- Errores en los datos
  p.ej. errores de transcripción en atributos simbólicos.
- Errores de redondeo
  p.ej. precisión finita de las medidas
  y de la representación de los datos en coma flotante.
- Valoraciones imprecisas
  p.ej. valores intervalares y difusos.

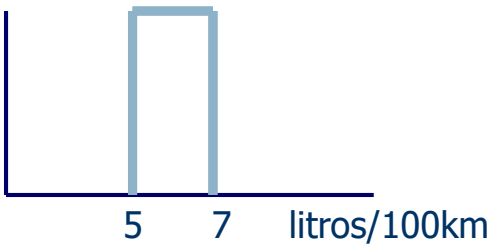
Limpieza de datos

Valores imprecisos

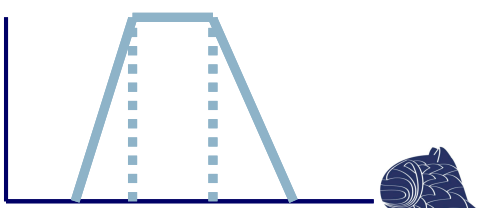
- Valor preciso:
  Consumo 6 l/100km



- Valor intervalar:
  Consumo entre 5 y 7 l/100km



- Valor difuso:
  Consumo moderado



Limpieza de datos

Datos incompletos

A menudo, los atributos con los que trabajamos con "asimétricos": sólo se considera importante su presencia.

EJEMPLOS

- Palabras de un documento
- Artículos incluidos en una transacción

En un supermercado, ¿le diríamos a un amigo algo como lo siguiente? "Veo que mi compra es muy similar a la tuya, ya que coincidimos en no comprar la mayor parte de los productos"…

Tratamiento de valores nulos

En ocasiones, nos pueden faltar datos:

- Falta de datos: información sesgada, datos dispersos…
- Valores nulos en los atributos
  (falta el valor del atributo, ya sea porque lo desconocemos o porque no resulta aplicable).

En el primer caso, es difícil asegurar la calidad de los datos: una gran cantidad no asegura su calidad (GIGO).

En el segundo caso, se puede hacer algo…

Tratamiento de valores nulos

Origen de los datos perdidos

- Valor no recogido
  (desconocimiento del valor sin factores aleatorios).
- Propiedad no aplicable
  (no se puede identificar el valor nulo con un 0/NO).
- Error de origen aleatorio
  - Totalmente aleatorio
    [MCAR: missing completely at random]
  - Aleatorio condicionado
    [MAR: missing at random]

Tratamiento de valores nulos

Estrategias disponibles

Eliminación de datos perdidos

- En situaciones totalmente aleatorias [MCAR]
- Cuando el volumen de datos disponible
  no queda seriamente afectado.

Imputación de datos

- Se estima el valor del dato perdido y "se rellena" el conjunto de datos (p.ej. valor más frecuente, media [MCAR], media condicionada [MAR], interpolación…)

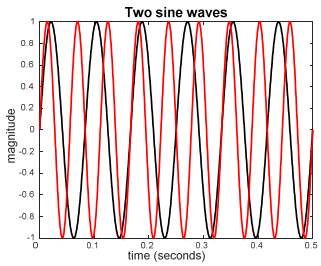
Uso de valores nulos

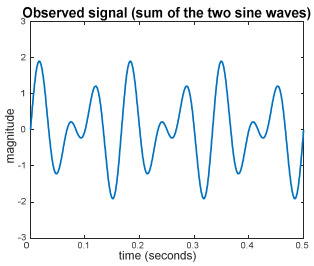
- Se crea un nuevo valor (p.ej. NS/NC, NA), que deberemos tratar adecuadamente durante el análisis.

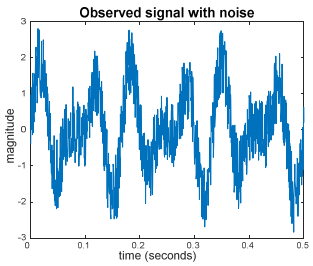
Presencia de ruido

Ruido

Modificación de los valores originales de los atributos







- **Ruido aleatorio**:
  medidas de sensores, errores de transcripción…
- **"Ruido" no aleatorio** (accidental o intencionado):
  Valores sin sentido, p.ej. dimensiones negativas.
  Valores por defecto, i.e. valores nulos "disfrazados", p.ej. 1 de enero como fecha de nacimiento.

Presencia de ruido

Estrategias disponibles

Depuración de los datos [data scrubbing]

Conocimiento específico para detectar y corregir errores.
p.ej. Códigos postales, correctores ortográficos…

Suavizado de los datos [smoothing]

- **Binning**: Se ordenan y dividen los datos en intervalos de la misma frecuencia, tras lo cual se suavizan con la media, la mediana o los límites de los intervalos.
- **Regresión**: Se ajustan los datos a una función.

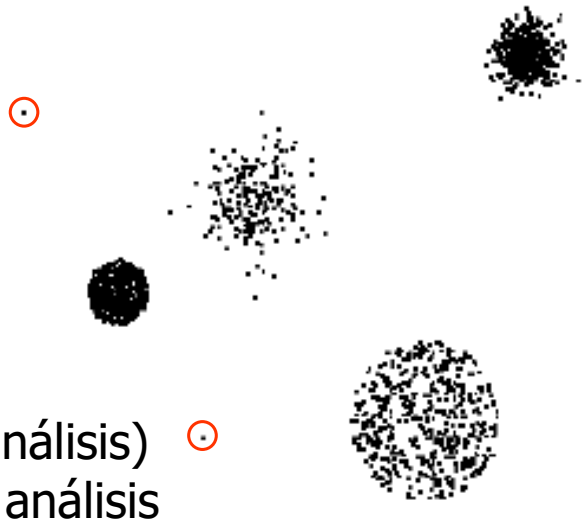
Agrupamiento de los datos [clustering]

permite detectar patrones y eliminar anomalías [outliers]

Presencia de ruido

Anomalías [outliers]

Datos con características considerablemente diferentes a la mayoría del resto de datos de nuestro conjunto de datos.



Pueden ser ruido
(interfiere con nuestro análisis)
o el objetivo de nuestro análisis
(detección de anomalías).

Transformación de datos

Antes de poder aplicar algunas técnicas de minería de datos, en la mayoría de los casos es necesario transformar los datos:

- Rangos de los valores de las distintas variables
  p.ej. normalización y estandarización/tipificación
- Cambios de tipos de datos
  p.ej. discretización (continuo → categórico ordinal)
- Reducción del volumen de datos
  p.ej. agregación/resumen, muestreo…
- Reducción del número de variables/dimensiones
  p.ej. selección y extracción de características

Normalización y estandarización

Técnicas de cambio de escala
(cuando existen diferencias excesivas en los rangos de las diferentes variables de nuestro conjunto de datos)

- Normalización [min-max normalization]
- Tipificación/estandarización [z-score normalization]
- Tipificación robusta
- Normalización decimal [decimal scaling]

Normalización y estandarización

Técnicas de cambio de escala
(cuando existen diferencias excesivas en los rangos de las diferentes variables de nuestro conjunto de datos)

- **Normalización** [min-max normalization]
  Habitualmente, normalización 0-1:

$$v' = \frac{v - min}{max - min}$$

- **Tipificación/estandarización**
  [**z-score** normalization]
  teniendo en cuenta media y desviación: $v' = \frac{v - \mu}{\sigma}$

Normalización y estandarización

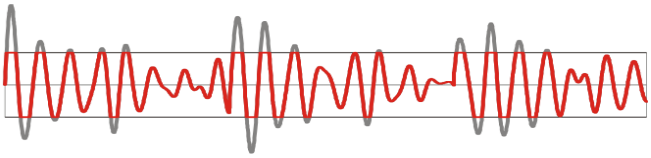
- **Tipificación robusta**

$$v' = \frac{v - mediana}{rango\ intercuartil}$$

- **Normalización por escalado decimal** [normalization by decimal scaling]
  donde s es el menor entero tal que max(|v'|)<1

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

Extra:



Winsorización

Recorte de outliers más allá de un percentil especificado, p={5,10},
p.ej. Clipping por debajo del percentil 5 y por encima del 95

Discretización

Sustitución de atributos numéricos
por atributos categóricos (ordinales):

- Se divide el rango de un atributo continuo en intervalos.
- Las etiquetas de los intervalos sustituyen a los datos originales.
- Se puede considerar información adicional además de los valores del atributo continuo (p.ej. discretización supervisada)

Discretización

Métodos de discretización

Discretización no supervisada

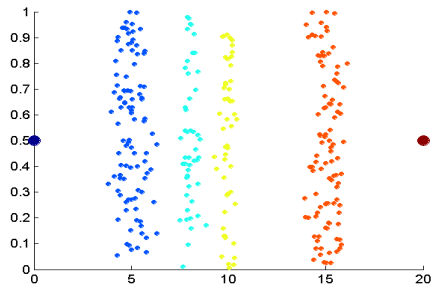
- División en intervalos [binning]:
  equi-width (distancia) vs. equi-depth (frecuencia)
- Análisis del histograma.
- Métodos de agrupamiento.

Discretización supervisada

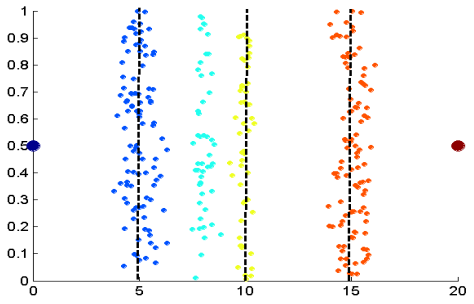
- Análisis de correlaciones
  (p.ej. Chi-merge χ^2).
- En la construcción de clasificacdores
  (p.ej. árboles de decisión).

Discretización

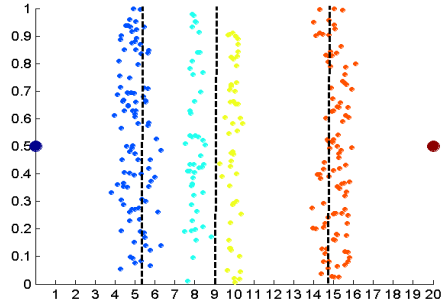
Datos originales



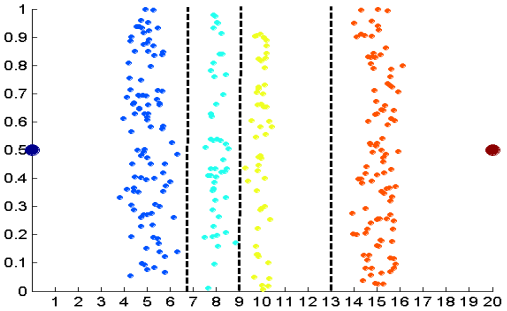
Equi-width (distancia)



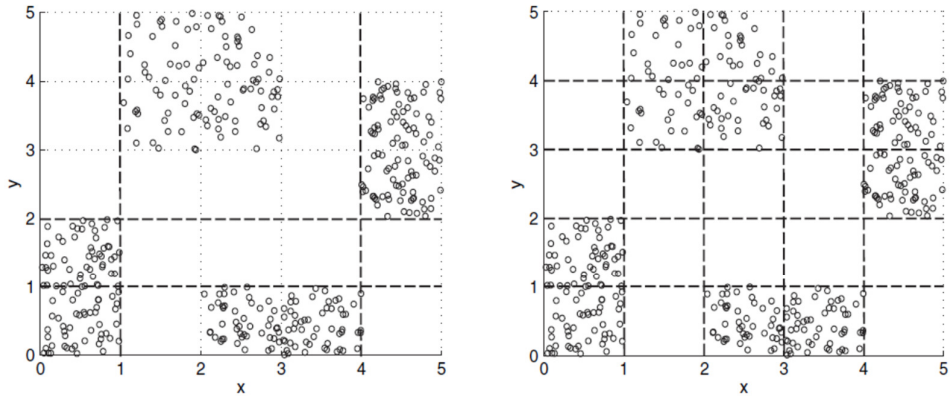
Equi-depth (frecuencia)



Agrupamiento (k-means)



Discretización supervisada

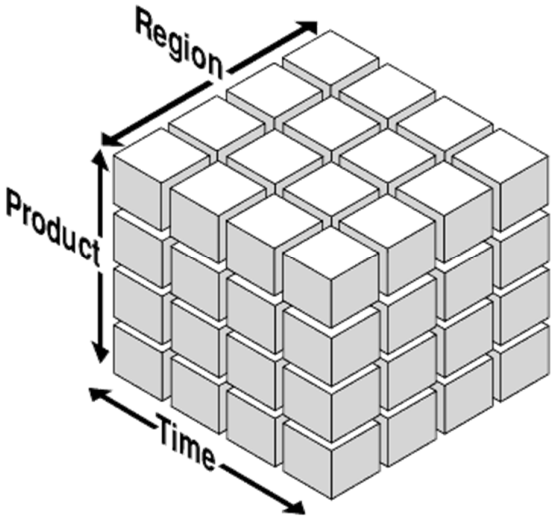


(a) Three intervals (b) Five intervals

Figure 2.14. Discretizing x and y attributes for four groups (classes) of points.

Reducción de datos

En ocasiones, disponemos de datos demasiado detallados y, por problemas de escalabilidad, no podemos utilizar la técnica de minería de datos deseada sobre el conjunto de datos original, p.ej. OLAP



Reducción de datos

Métodos de reducción de datos
[a.k.a. data size reduction / numerosity reduction]

Métodos paramétricos, p.ej. regresión
Se asume que los datos se ajustan a un modelo, se estiman los parámetros del modelo y se descartan los datos (excepto, tal vez, posibles outliers).

Métodos no paramétricos

- Agregación de datos en cubos de datos (usando jerarquías de conceptos)
- Técnicas de muestreo

Técnicas de muestreo

IDEA: Obtener una muestra pequeña s
que represente al conjunto de datos completo N ($s \ll N$).

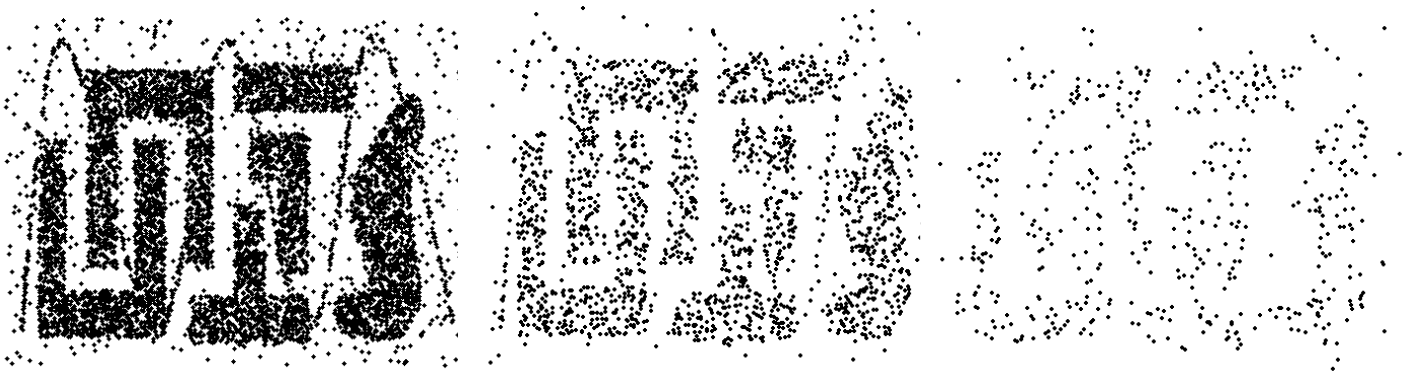
- Se puede conseguir un algoritmo de minería de datos con una **eficiencia sublineal** con respecto al tamaño del conjunto de datos original.

La clave es que la muestra sea representativa:

- Una muestra aleatoria puede no resultar adecuada ante la presencia de sesgos en los datos.
- Se pueden idear métodos de muestreo adaptativo, p.ej. muestreo estratificado

Técnicas de muestreo

Muestreo aleatorio

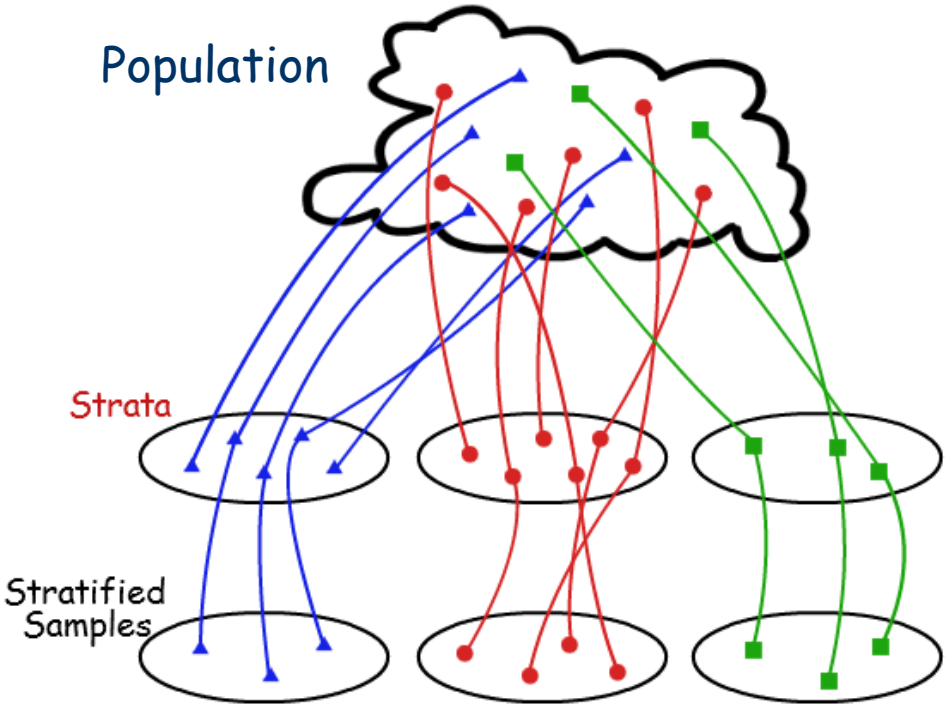


8000 points | 2000 Points | 500 Points

Si la muestra no es representativa (p.ej. es demasiado pequeña), se pierden algunas propiedades de interés del conjunto de datos original :-(

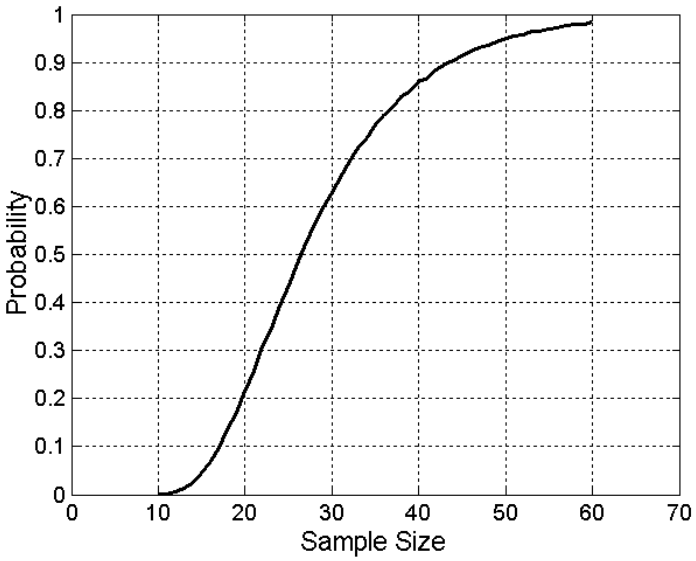
Técnicas de muestreo

Muestreo estratificado



Técnicas de muestreo

Tamaño muestral



Tamaño de la muestra necesario para obtener, al menos, un ejemplo de cada uno de 10 grupos del mismo tamaño (p.ej. 10 clases igualmente frecuentes).

Reducción de dimensionalidad

La maldición de la dimensionalidad

- Conforme aumenta el número de dimensiones, los datos están más dispersos (lejos unos de otros).
- El número de subespacios crece exponencialmente con el número de dimensiones.

Las **técnicas de reducción de la dimensionalidad** reducen el número de variables/dimensiones consideradas:

- Para evitar la maldición de la dimensionalidad.
- Para eliminar características irrelevantes y reducir ruido.
- Para reducir los recursos necesarios (tiempo & espacio).
- Para facilitar la visualización de los datos.

Reducción de dimensionalidad

Métodos de reducción de la dimensionalidad

Selección de características

Encontrar un subconjunto adecuado de las variables, características o atributos originales.

Extracción de características

Construcción de nuevas variables/características a partir de los datos disponibles con el objetivo de reducir su dimensionalidad.

p.ej. Análisis de componentes principales (PCA)

Extracción de características

Análisis de componentes principales

[PCA: Principal Component Analysis]

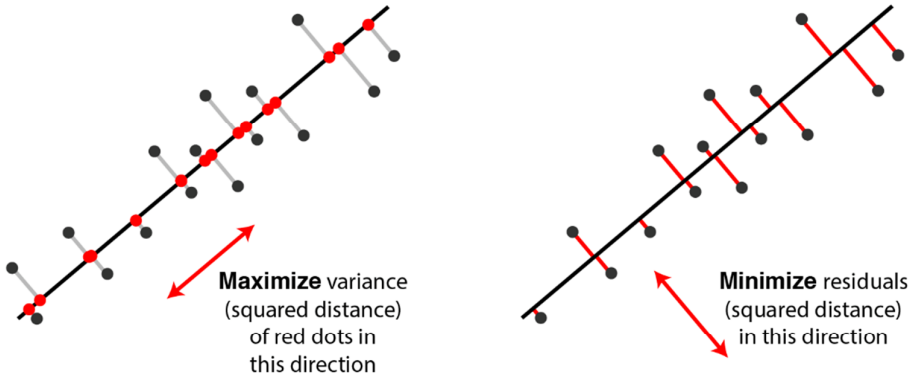
Técnica estadística que usa una transformación ortogonal para convertir un conjunto de observaciones de variables posiblemente correladas en un conjunto de valores de variables linealmente no correladas (denominadas componentes principales).

ALGORITMO: Usar los vectores propios o autovectores [eigenvectors] de la matriz de covarianza para definir el espacio de componentes principales.

Extracción de características

Análisis de componentes principales

[PCA: Principal Component Analysis]



Two equivalent views of principal component analysis.

Extracción de características

Análisis de componentes principales

[PCA: Principal Component Analysis]

- Covarianza

$$\mathbf{C} = \frac{1}{n-1}\left((X-\bar{x})^\top (X-\bar{x})\right) \qquad \mathbf{C} = \frac{1}{n-1} X^\top X$$ X ya centrado

- Varianza de la proyección Xw

$$\frac{1}{n-1}(\mathbf{X}\mathbf{w})^\top \mathbf{X}\mathbf{w} = \mathbf{w}^\top \left(\frac{1}{n-1}\mathbf{X}^\top \mathbf{X}\right)\mathbf{w} = \mathbf{w}^\top \mathbf{C}\mathbf{w}$$

- Problema de optimización

$$\begin{aligned} &\underset{\mathbf{w}}{\text{maximize}} && \mathbf{w}^\top \mathbf{C}\mathbf{w} \\ &\text{subject to} && \mathbf{w}^\top \mathbf{w} = 1 \end{aligned}$$

$$L = \mathbf{w}^\top \mathbf{C}\mathbf{w} - \lambda(\mathbf{w}^\top \mathbf{w} - 1)$$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{C}\mathbf{w} - \lambda\mathbf{w} = 0$$

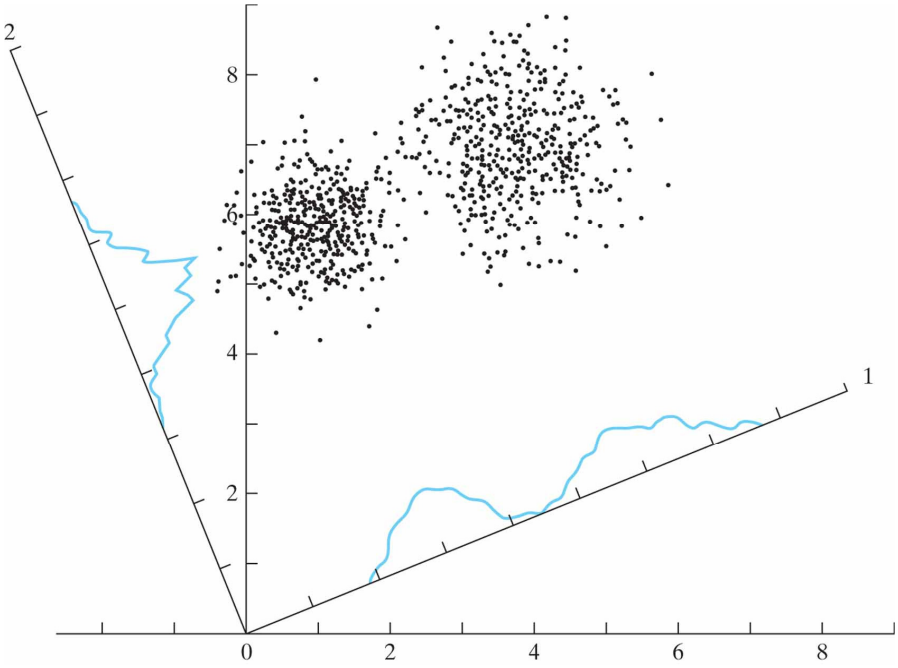
$$= \mathbf{C}\mathbf{w} = \lambda\mathbf{w}$$

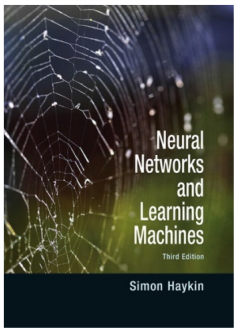
Eigendecomposition

Extracción de características

Análisis de componentes principales

[PCA: Principal Component Analysis]

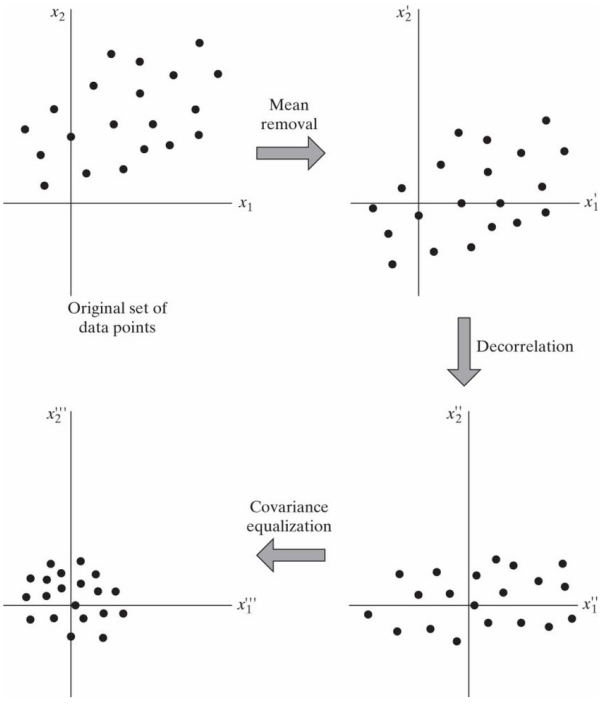


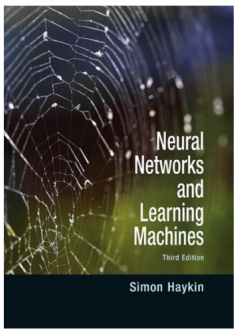


Extracción de características

Análisis de componentes principales

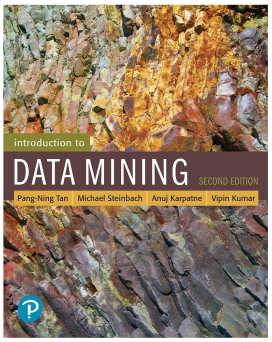
[PCA: Principal Component Analysis]





Extracción de características

Análisis de componentes principales



Extracción de características

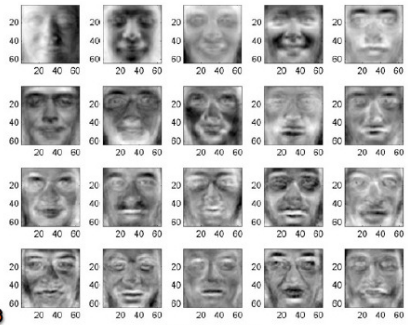
Análisis de componentes principales

[PCA: Principal Component Analysis]

64x64 images of faces = 4096 dimensional data

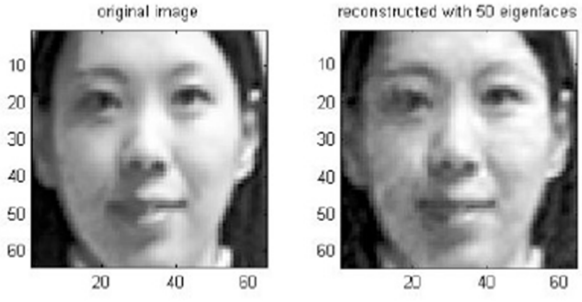
Average Face

+



Principal Components

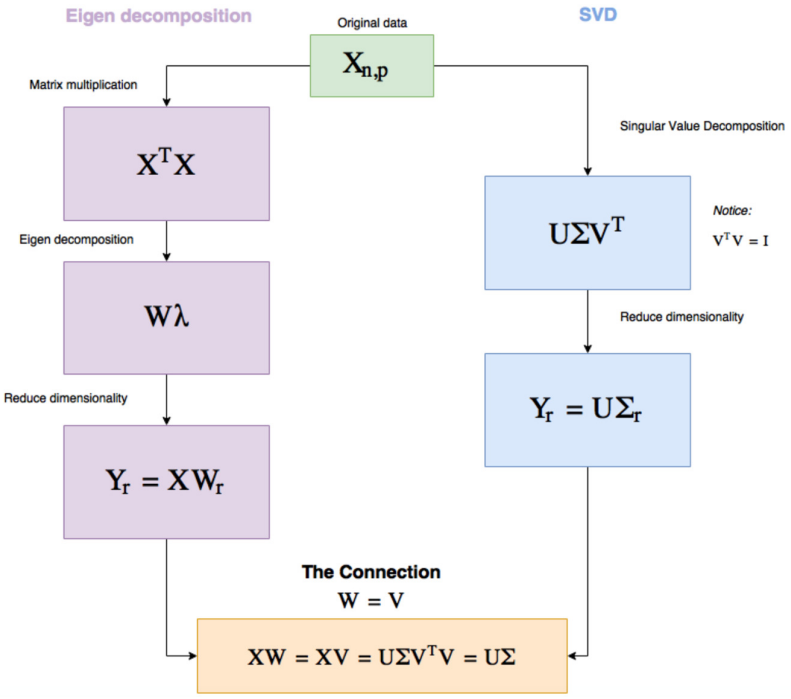
Eigenfaces



Extracción de características

Análisis de componentes principales

PCA [Principal Component Analysis] vs. SVD [Singular Value Decomposition]



Extracción de características

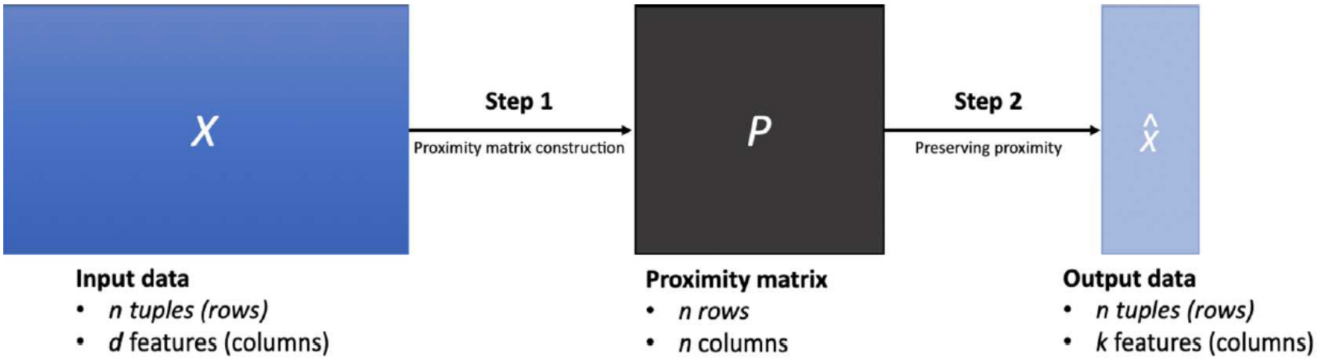
Métodos no lineales de reducción de la dimensionalidad

PCA es un método lineal (cada componente principal es una combinación lineal de las variables originales) que funciona bien si los datos siguen una distribución normal o forman clusters linealmente separables.

Cuando los datos no son linealmente separables, se construye una matriz de proximidad P y se aprende una matriz con k dimensiones (k≪d) que preserve esa proximidad…

Extracción de características

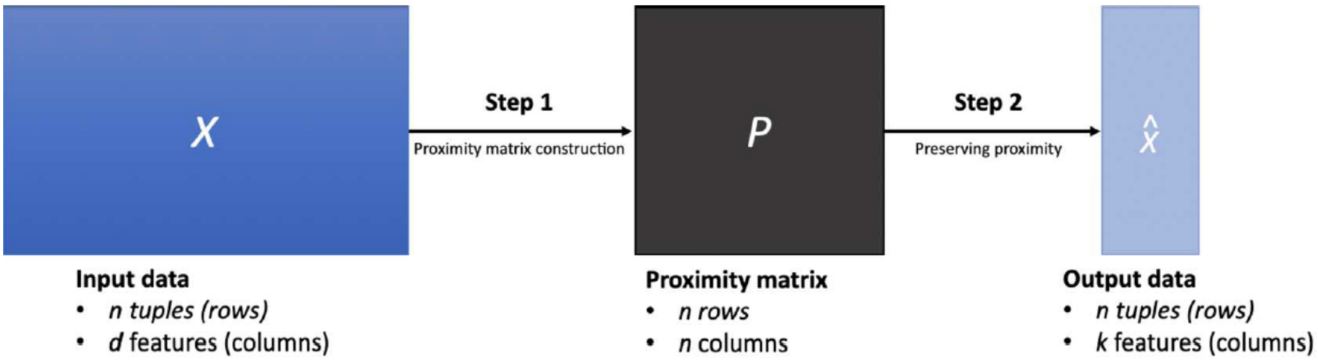
Métodos no lineales de reducción de la dimensionalidad



- **KPCA [Kernel PCA]**
- **SNE [Stochastic Neighborhood Embedding]**

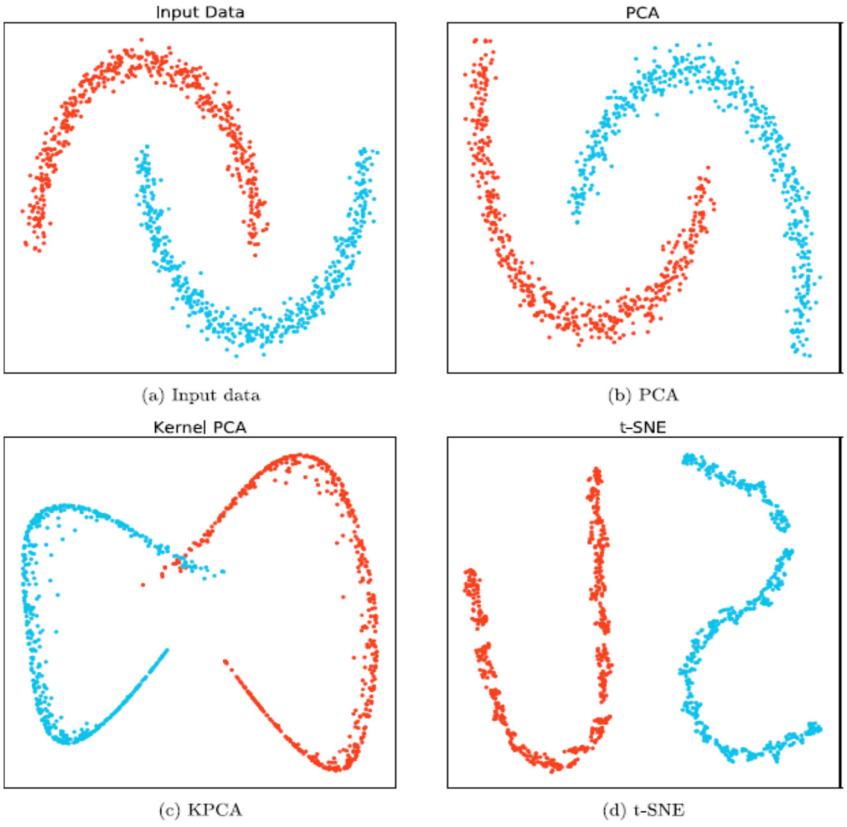
Extracción de características

Métodos no lineales de reducción de la dimensionalidad



	Step 1: Proximity Construction	Step 2: Preserving Proximity
KPCA	$P(i,j) = \kappa(x_i, x_j)$	$\min \sum_{i,j=1}^{n} (P(i,j) - \hat{P}(i,j))^2 = \|P - \hat{P}\|_{fro}^2$
SNE	$P(i,j) = \frac{e^{-d_{ij}^2}}{\sum_{l=1, l\neq i}^{n} e^{-d_{il}^2}}$	$\min \sum_{i=1}^{n} \mathrm{KL}(P_i \| \hat{P}_i)$

Extracción de características



(a) Input data (b) PCA

(c) KPCA (d) t-SNE

Selección de características

Hay 2^d combinaciones de atributos si nuestros datos tienen d dimensiones/variables/atributos/características.

No se puede realizar una exploración exhaustiva de ese espacio de búsqueda, por lo que se emplean criterios heurísticos:

- Mejor atributo individual asumiendo independencia entre los diferentes atributos (test estadístico).
- Algoritmo greedy de selección de atributos.
- Algoritmo greedy de eliminación de atributos

Selección de características

Estrategias de selección de características

Selección incluida [embedded]

El propio algoritmo de minería de datos va seleccionando la variable más adecuada (p.ej. árboles de decisión).

Filtrado de variables [filter]

Se seleccionan las variables antes de aplicar las técnicas de minería de datos (p.ej. términos en text mining).

Selección por cobertura [wrapper]

La bondad del resultado de aplicar la técnica de minería de datos sirve como criterio de selección.

Ingeniería de características

a.k.a. feature engineering

- Crear nuevos atributos que capturen información importante para un problema.
- El uso de técnicas de minería de datos, con las características "adecuadas", puede resultar mucho más eficiente.

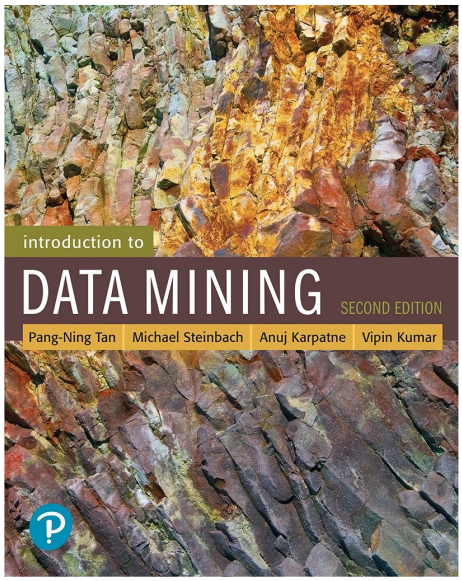
Ingeniería de características

Estrategias

- Construcción de características
  p.ej. densidad = masa / volumen
- Extracción de características
  p.ej. fronteras en imágenes [edge detection]
- Transformación a un nuevo espacio
  p.ej. análisis de Fourier (frecuencias), wavelets…
- Aprendizaje de representaciones (automático)
  p.ej. deep learning

Bibliografía

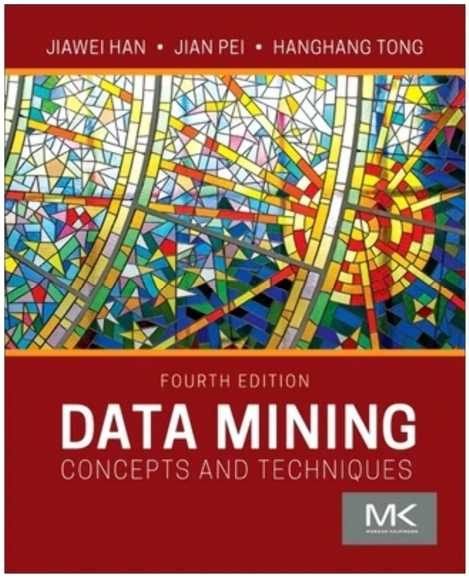
Pang-Ning Tan,
Michael Steinbach,
Vipin Kumar &
Anuj Karpatne:
Introduction to Data Mining,
2nd edition, Addison Wesley, 2018.
ISBN 0133128903



2.1 Types of Data
2.2 Data Quality
2.3 Data Preprocessing
2.4 Measures of Similarity and Dissimilarity

Bibliografía

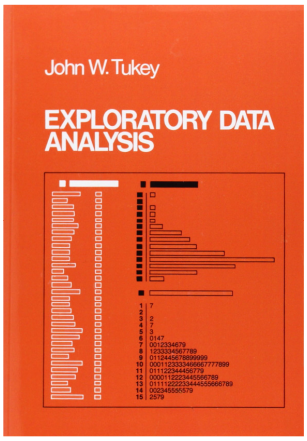
Jiawei Han,
Jian Pei &
Hanghang Tong:
Data Mining:
Concepts and Techniques,
4th edition, Morgan Kaufmann, 2022.
ISBN 0128117605



2 Data, measurements, and data preprocessing
3 Data warehousing and online analytical processing

Bibliografía complementaria

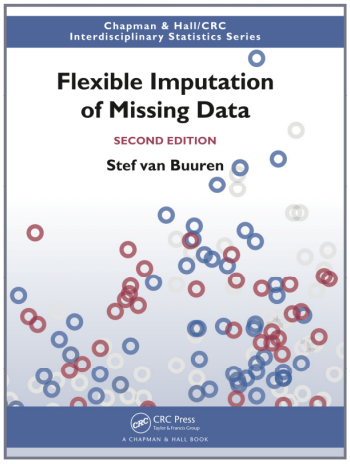
- John Tukey:
  Exploratory Data Analysis,
  Addison-Wesley, 1977.
  ISBN 0201076160.



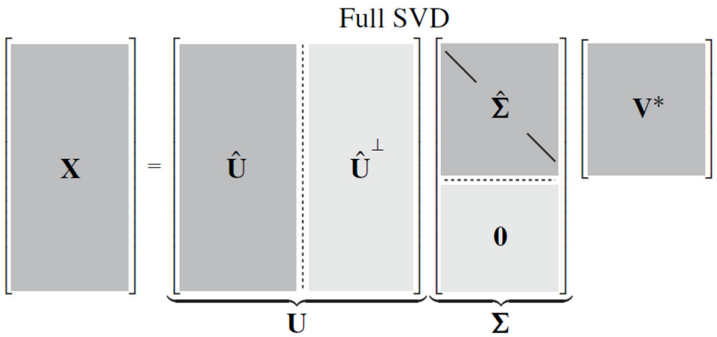
- **NIST/SEMATECH**
  e-Handbook of Statistical Methods,
  http://www.itl.nist.gov/div898/handbook
  1. Exploratory Data Analysis
  https://www.itl.nist.gov/div898/handbook/eda/eda.htm

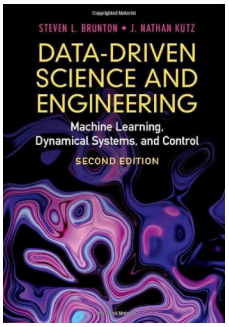
Bibliografía complementaria

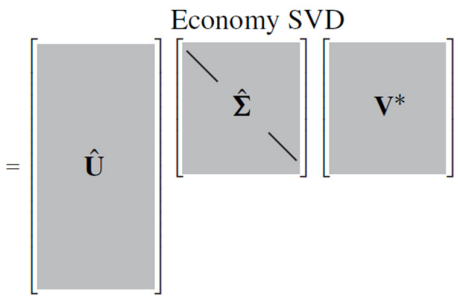
- Stef van Buuren:
  Flexible Imputation of Missing Data
  2nd edition, CRC Press, 2018.
  ISBN 1138588318
  https://stefvanbuuren.name/fimd/



Apéndice
SVD [Singular Value Decomposition]



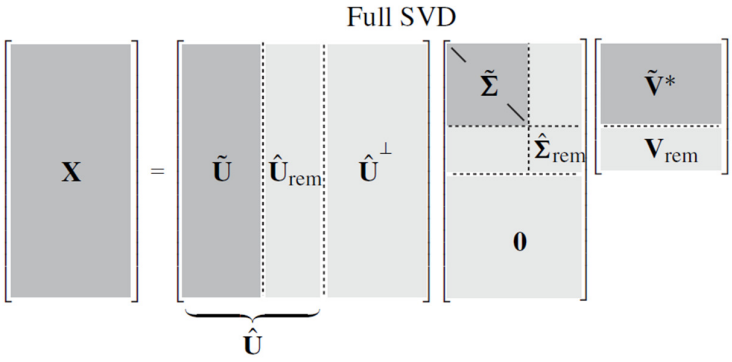


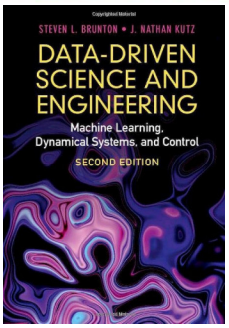


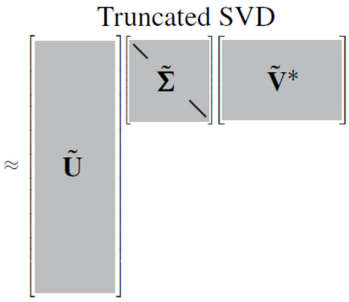
Python

```
import numpy as np
# Full SVD
U, S, VT = np.linalg.svd(X,full_matrices=True)
# Economy SVD
Uhat, Shat, VThat = np.linalg.svd(X,full_matrices=False)
```

Apéndice
SVD [Singular Value Decomposition]



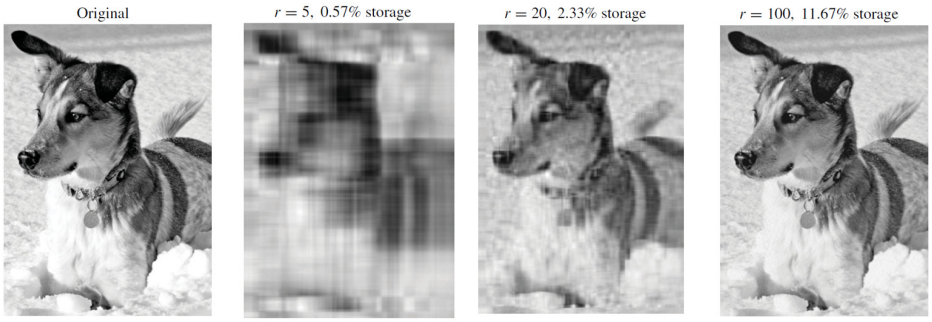


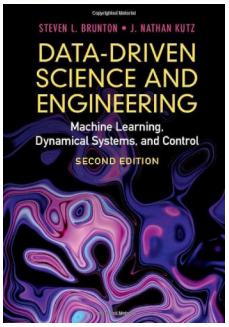


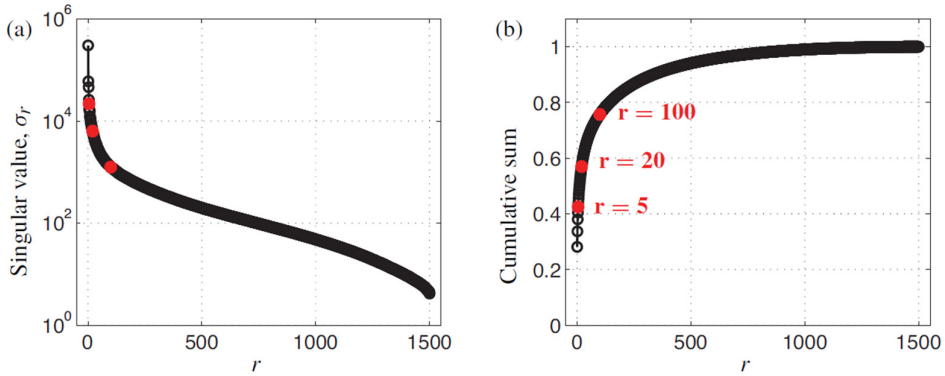
Teorema de Eckart-Young

$$\underset{\tilde{\mathbf{X}},\ \text{s.t. rank}(\tilde{\mathbf{X}})=r}{\operatorname{argmin}} \quad \|\mathbf{X}-\tilde{\mathbf{X}}\|_F = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^*$$

Apéndice
SVD [Singular Value Decomposition]







Compresión de imágenes con SVD