

UNIVERSIDAD  
**AUSTRAL**



**# SOMOSAUSTRAL**

UNIVERSIDAD  
**AUSTRAL**



INGENIERÍA

---

**# SOMOSAUSTRAL**

# Introduccion a DM

Datos y Preprocesamiento – 2022

Mg. Lic. Gastón Pezzuchi, MSc.

# Datos y Preprocesamiento

Introduction to Data Mining (2nd. Edition)

# Conozca sus Datos



INGENIERÍA

---

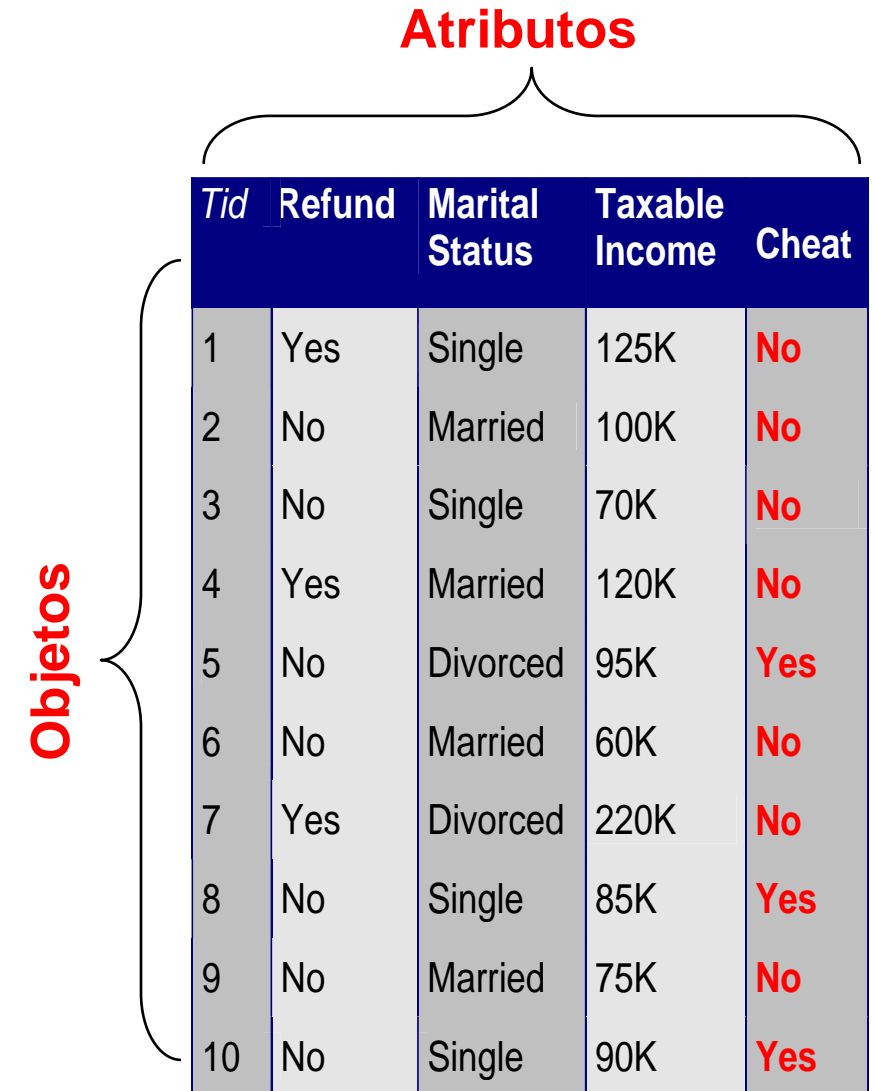
# Agenda

---

- Atributos y Objetos
- Tipos de Datos
- Calidad de los datos
- Similaridad y Distancia
- Pre-procesamiento de datos

# Que son los datos?

- ❑ Colección de objetos dato (**data objects**) y sus **atributos**
- ❑ Un **atributo** es una propiedad o característica de un objeto que puede variar de objeto en objeto o a lo largo del tiempo.
  - Ej.: color de ojos de una persona, temperatura, etc.
  - Atributo también se conoce como variable, campo, característica, dimensión o *feature*
- ❑ Una colección de atributos describe a un **objeto**
  - Objeto también se conoce como registro, punto, caso, entidad o instancia



A diagram illustrating the relationship between objects and attributes. A table with 10 rows and 5 columns is shown. A vertical curly brace on the left side of the table is labeled 'Objetos' in red. A horizontal curly brace at the top of the table is labeled 'Atributos' in red.

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

# Una mirada mas completa sobre los datos

---

- Los datos pueden tener partes
- Las diferentes partes de los datos pueden tener relaciones
- Mas generalmente, los datos pueden tener estructura
- Los datos pueden ser incompletos..
- ...



# Valores de los Atributos

---

- Los **Valores de los atributos** son números o símbolos asignados a un atributo de un objeto en particular
  
- Distinción entre atributos y valores de los atributos
  - El mismo atributo puede ser mapeado a diferentes valores
    - ◆ Ej.: la altura puede ser medida en pies o metros
  
  - Diferentes atributos pueden ser mapeados al mismo conjunto de valores
    - ◆ Ej.: Los valores de los atributos para ID y edad son enteros
    - ◆ PERO las propiedades de los valores de los atributos pueden ser diferentes de las propiedades de los atributos...

# Medición

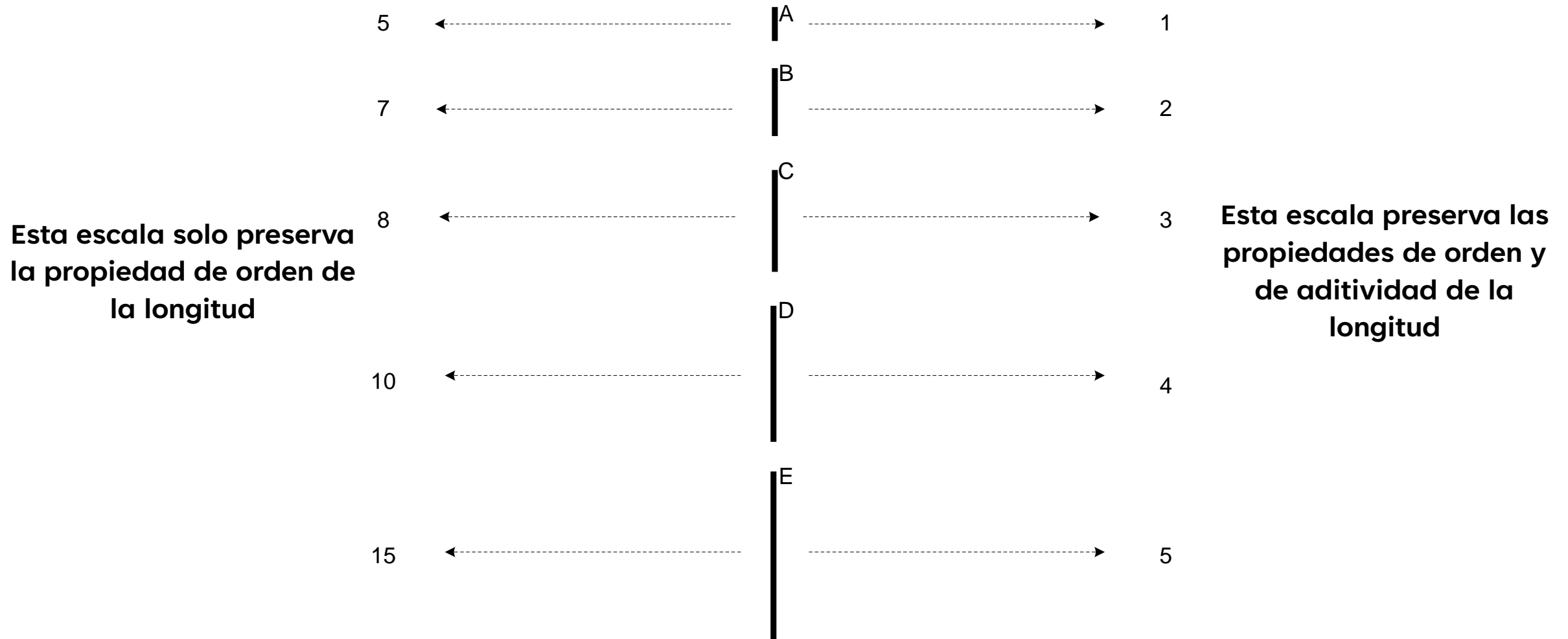
---

- Una **escala de medición** es una regla (función) que asocia un valor numérico o simbólico con un atributo de un objeto
- El **proceso de medición** es la aplicación de una escala de medición para asociar un valor con un atributo particular de un objeto específico.
- Es usual referirnos a los tipos de los atributos como los **tipos de escalas de medición**.

# Ej. Medición de la Longitud

---

- La forma en la que se mide un atributo puede no *matchear* las propiedades de los atributos.



# Tipos de Atributos

---

- Hay diferentes tipos de atributos (Stevens, 1946)
  - Nominal
    - ◆ Ej.: ID, color de ojo, código postal
  - Ordinal
    - ◆ Ej.: rankings (ej., sabor de una comida en una escala de 1-10), grados, altura {alto, medio, bajo}
  - Intervalo
    - ◆ Ej.: temperaturas en Celsius o Fahrenheit.
  - Razón
    - ◆ Ej.: temperatura en Kelvin, longitud, tiempo, conteos

[Stevens, S. S.](#) (7 June 1946). "On the Theory of Scales of Measurement". [Science](#). **103** (2684): 677–680. doi:[10.1126/science.103.2684.677](https://doi.org/10.1126/science.103.2684.677).

# Propiedades de los Valores de los Atributos

---

□ De acuerdo al tipo de atributo, son las propiedades/operaciones disponibles:

- Distintividad:  $= \neq$
- Orden:  $< >$
- Suma:  $+ -$  (cálculo de diferencias)
- Multiplicación:  $* /$  (cálculo de cocientes)
- Atributos nominales: distintividad
- Atributos Ordinales: distintividad & orden
- Atributos Intervalares: distintividad, orden & diferencias
- Atributos de Razón: las 4 propiedades / operaciones

# Diferencias entre los niveles de Razón y de Intervalo

---

- Tiene sentido físico decir que una temperatura de  $10^{\circ}$  es el doble que una de  $5^{\circ}$  en
  - la escala Celsius?
  - la escala Fahrenheit?
  - la escala Kelvin?
  
- Supongamos medir la altura por encima del promedio
  - Si la altura de Pedro es 3 centímetros por encima del promedio y la de José es 6 centímetros por encima del promedio, podemos decir que José es el doble de alto que Pedro?
  - Es esta situación análoga a la de la temperatura?

		Attribute Type	Description	Examples	Operations
Categorical Qualitative		Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: { <i>male</i> , <i>female</i> }	mode, entropy, contingency correlation, $\chi^2$ test
		Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, { <i>good</i> , <i>better</i> , <i>best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative		Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
		Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

**Categorización de S. S. Stevens** (Stevens, S.S. (1946). On the theory of Scales Measurement. *Science*, 103(2684), 677-680.)

		Attribute Type	Transformation	Comments
Categorical Qualitative		Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
		Ordinal	An order preserving change of values, i.e., $new\_value = f(old\_value)$ where $f$ is a monotonic function	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Numeric Quantitative		Interval	$new\_value = a * old\_value + b$ where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
		Ratio	$new\_value = a * old\_value$	Length can be measured in meters or feet.



# Atributos Discretos y Continuos

---

## □ Atributos Discretos

- Conjunto finito o infinito numerable (contable) de valores
- Ej.: códigos postales, conteos, o conjuntos de palabras en una colección de documentos
- Muchas veces representados por valores enteros.
- Nota: los **atributos binarios** son un caso especial de los atributos discretos

## □ Atributos Continuos

- Tiene números reales como valor del atributo
- Ej.: temperatura, altura o peso.
- En la practica, los valores reales pueden ser medidos y representados empleando un numero finito de dígitos.
- Los atributos continuos se representan típicamente como variables de coma flotante.

# Atributos Asimétricos

---

- Solo la presencia (un valor no nulo del atributo) se considera importante.
  - ◆ Palabras presentes en documentos
  - ◆ Ítems presentes en transacciones de clientes
  
- Si nos encontramos con un amigo en un almacén, alguna vez diríamos algo como?:  
*“Veo que nuestras compras son muy **similares** dado que no compramos la mayoría de las mismas cosas.”*
  
- Necesitamos dos atributos binarios asimétricos para representar un atributo binario ordinario
  - El análisis de asociaciones (*Association analysis*) emplea atributos asimétricos

# Extensiones y Criticas

---

- ❑ Velleman, Paul F., and Leland Wilkinson. "Nominal, ordinal, interval, and ratio typologies are misleading." *The American Statistician* 47, no. 1 (1993): 65-72.
- ❑ Mosteller, Frederick, and John W. Tukey. "Data analysis and regression. A second course in statistics." *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, Reading, Mass.: Addison-Wesley, 1977.
- ❑ Chrisman, Nicholas R. "Rethinking levels of measurement for cartography." *Cartography and Geographic Information Systems* 25, no. 4 (1998): 231-242.

# Criticas

---

## □ Incompletitud

- Binarios Asimétricos
- Cíclicos
- Multivariados
- Parcialmente ordenados
- Membresía parcial
- Relaciones entre los datos

## □ Los datos reales son aproximados y ruidosos

- Esto hace complicado el reconocimiento del tipo adecuado de atributo
- Tratar un tipo de atributo como si fuera otro, puede ser aproximadamente correcto.

# Criticas ...

---

- No es una buena guía para realizar análisis estadísticos
  - Puede restringir innecesariamente operaciones y resultados
    - ◆ El análisis estadístico es muchas veces aproximado
    - ◆ Ej. Utilizar análisis intervalar para valores ordinales.
  - Las transformaciones son usuales, pero no preservan escalas
    - ◆ Pueden transformar datos a una nueva escala con mejores propiedades estadísticas.
    - ◆ Muchos análisis estadísticos dependen solamente de la distribución.

# Ejemplos mas complicados

---

- IDs
  - Nominales, ordinales, o intervalares?
  
- Cantidad de cilindros en un motor de un auto
  - Nominal, ordinal, o razón?
  
- Escalas Sesgadas
  - Intervalos o Razón

# Importancia del Tipo de Atributo

---

- El tipo de operaciones que se elijan debe ser “significativa” para el tipo de dato con el que se cuenta
  - Distintividad, orden, intervalos significativos, y cocientes significativos son solo cuatro propiedades de los datos
  - El tipo de dato que uno observa (números o cadenas de texto) pueden no capturar todas las propiedades o sugerir propiedades que no están presentes
  - Los análisis pueden depender de esas otras propiedades de los datos
    - ◆ Muchos de los análisis estadísticos dependen solo de la distribución
  - Muchas veces, lo que es significativo (que tiene sentido) se mide en términos de significancia estadística
  - Pero en el fondo, lo que es significativo se mide en termino del dominio de aplicación.

# Comentarios finales sobre niveles de medicion

---

- Distintividad, orden, significado real de intervalos o de cocientes son solo cuatro propiedades posibles para los datos.
- La cantidad de números o símbolos empleados para capturar los valores de los atributos pueden no capturar todas las propiedades de los atributos o sugerir propiedades que no están presentes.
- Los datos muchas veces se transforman para ser analizados.
- La evaluación final en cualquier análisis de datos, incluyendo operaciones sobre los atributos es si los resultados tienen sentido desde el dominio de negocios.



# Tipos de conjuntos de datos

---

- Registros
  - Matriz de Datos
  - Documentos
  - Transacciones
- Grafos
  - *World Wide Web*
  - Estructuras Moleculares
- Datos ordenados
  - Datos Espaciales
  - Datos Temporales
  - Datos Secuenciales
  - Secuenciación Genética

# Características importantes de los datos

---

- Dimensionalidad (cantidad de atributos)
  - ◆ Problemas derivados de la alta dimensionalidad
- Dispersión (escasez)
  - ◆ Solo importa la presencia
- Resolución
  - ◆ Los patrones dependen de la escala
- Tamaño
  - ◆ El tipo de análisis puede depender del tamaño del conjunto de datos

# Registros

---

- Los datos están representados por una colección de registros, cada uno compuesto por un conjunto fijo de atributos.

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

# Matriz de datos

---

- Si los objetos tienen el mismo numero fijo de atributos numéricos, es posible pensar en cada uno como un punto en un espacio multidimensional, donde cada dimensión representa un atributo distinto.
- Estos datos pueden ser representados entonces por una matriz de  $m$  por  $n$ , donde hay  $m$  filas (una por objeto) y  $n$  columnas, una por cada atributo.

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

# Documentos

---

- Cada documento se convierte en un vector de ‘términos’
  - Cada termino es un componente (atributo) del vector
  - El valor de cada componente es la cantidad de veces que ese termino aparece en el documento.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

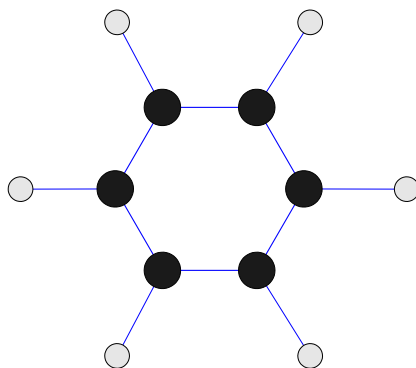
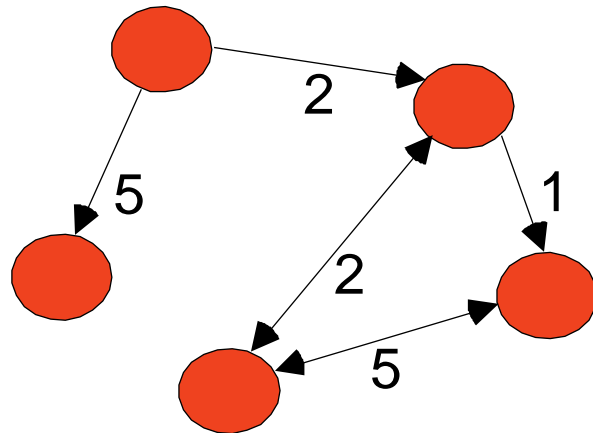
# Datos transaccionales

---

- Son un tipo especial de registro, donde
  - Cada registro (transacción) involucre un conjunto de ítems.
  - Por ejemplo en un supermercado. El conjunto de productos adquiridos por un cliente durante una compra constituye una transacción, en la que los productos adquiridos son los ítems.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

## □ Ejemplos: Grafo genérico, molécula, página web



Benzene Molecule: C<sub>6</sub>H<sub>6</sub>

### Useful Links:

- [Bibliography](#)
- Other Useful Web sites
  - [ACM SIGKDD](#)
  - [KDnuggets](#)
  - [The Data Mine](#)

### Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- [Books](#)
- [General Data Mining](#)

### Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.  
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support)", John Wiley & Sons, 1997.

### General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

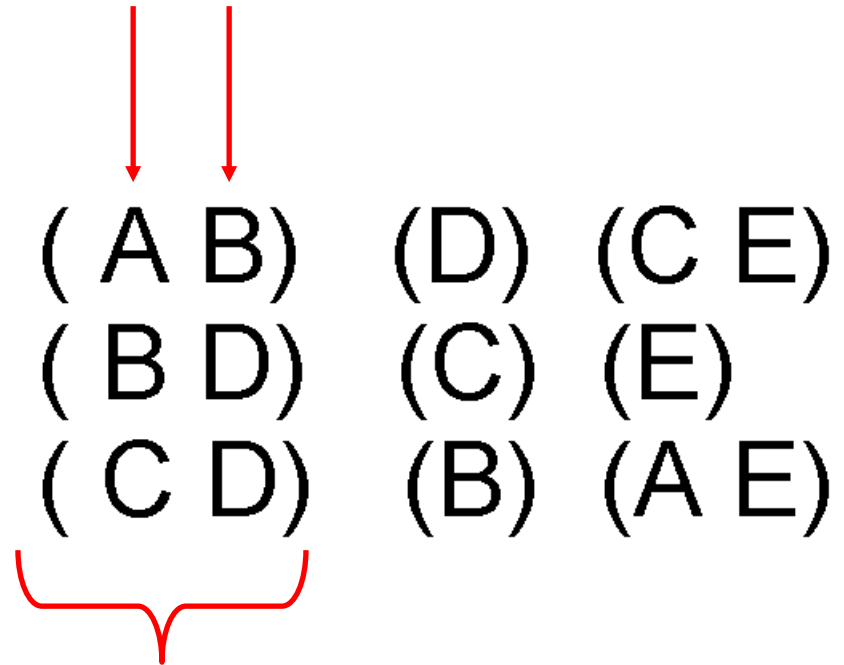
Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

# Datos ordenados

---

## □ Secuencias de transacciones

Ítems/Eventos



Un elemento de  
la secuencia



# Datos ordenados

---

## □ Datos de secuencia genómica

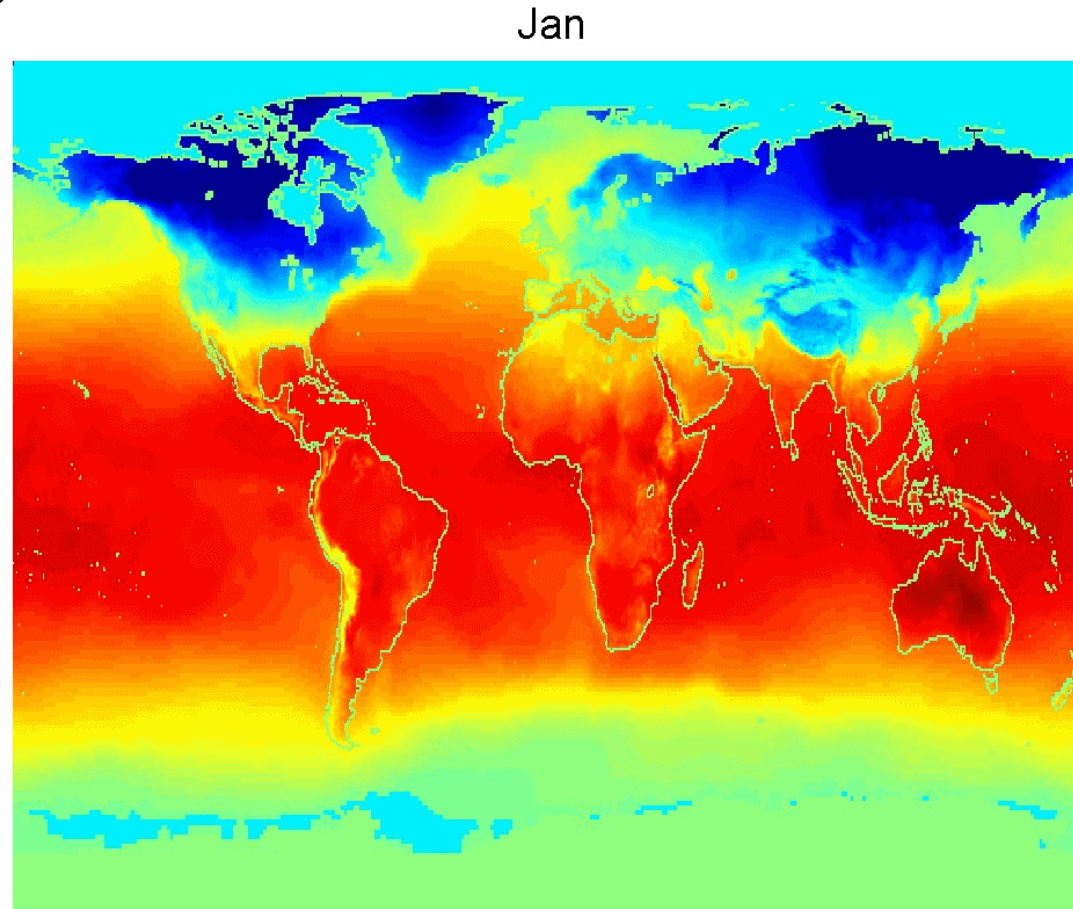
**GGTTC CGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG**

# Datos ordenados

---

## □ Datos Espacio-Temporales

**Temperatura mensual  
promedio de la tierra y  
el mar**



# Calidad del Datos – Recordar G.I.G.O

---

- Una mala calidad de los datos afecta en forma negativa cualquier esfuerzo de análisis.

*“The most important point is that poor data quality is an unfolding disaster.*

- Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate.”*

Thomas C. Redman, DM Review, August 2004

- Ej. DM: Un modelo de clasificación para detectar personas riesgosas para la obtención de un crédito se construye con datos de baja calidad
  - Se le deniega el crédito a algunas personas que podrían haberlo recibido
  - Se asigna una gran cantidad de créditos a personas que no los pagan.

# Calidad del dato...

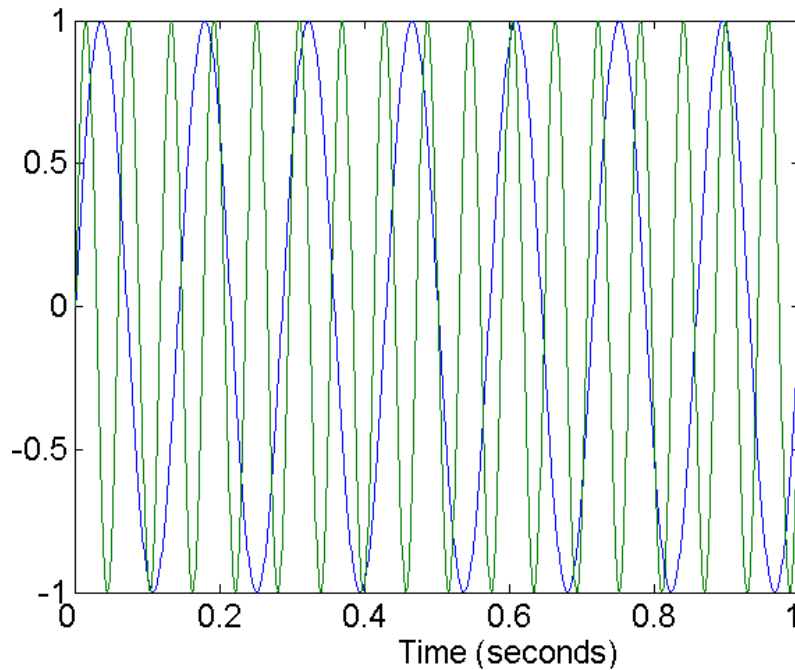
---

- Cuales son los tipos de problemas de calidad de datos?
- Como podemos detectar problemas con los datos?
- Que podemos hacer con estos problemas?
  
- Ejemplos:
  - Ruido y *outliers* (valores extremos)
  - *Missing values* (valores faltantes)
  - Datos duplicados
  - Datos incorrectos

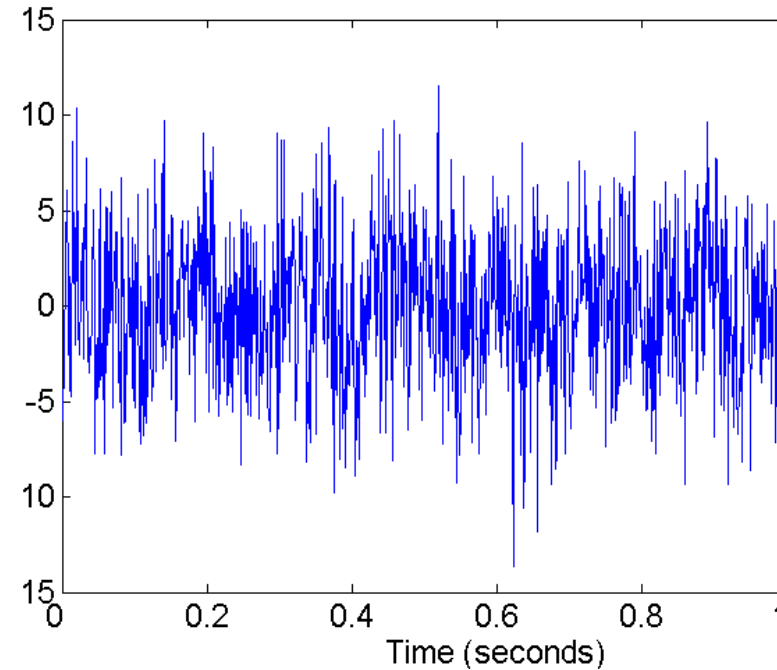
# Ruido

---

- Para los objetos, ruido es un objeto extraño
- Para atributos, ruido se refiere a la modificación de los valores originales
  - Ej: distorsión de la voz de una persona cuando habla en un teléfono con una conexión de baja calidad; “Lluvia” en la pantalla del televisor.



**Dos ondas senoidales**



**Dos ondas senoidales + Ruido**

# Ruido y Artefactos

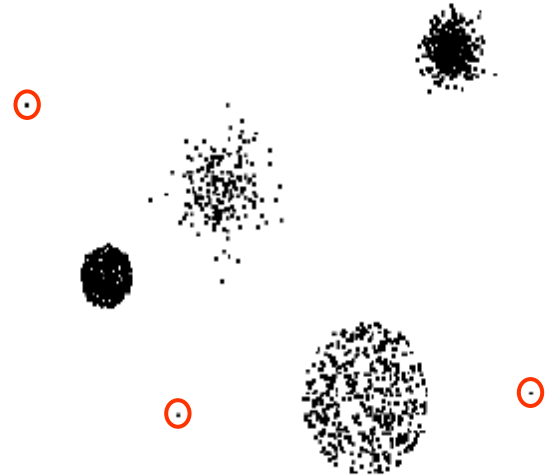
---

- El ruido es el componente aleatorio del error de medición.
  - Involucra la distorsión de un valor o el agregado de objetos espurios.
- Los errores de datos pueden resultar en un fenómeno mas determinístico, tal como una marca en el mismo lugar en un conjunto de fotografías, en este caso se determinan **artefactos**.

# Outliers (Valores extremos)

---

- **Outliers** son objetos de datos cuyas características son considerablemente diferentes que la mayoría de los otros objetos en el dataset
  - **Caso 1:** Los *outliers* son ruido que interfiere con el análisis de datos
  - **Case 2:** Los *outliers* son el objetivo de nuestro análisis
    - ◆ Fraude con tarjetas de crédito
    - ◆ Detección de Intrusos
- Causas?



# Missing Values (Valores Faltantes)

---

## □ Razón para los valores faltantes

- Información no relevada  
(ej., las personas no dan su edad y peso)
- Los atributos pueden no ser aplicables a todos los casos  
(ej., el ingreso anual puede no ser aplicable para niños)

## □ Como tratar los valores faltantes

- Eliminar los objetos o las variables
- Estimar los valores faltantes
  - ◆ Ej.: serie de tiempo de la temperatura
  - ◆ Ej.: datos censales
- Ignorar los valores faltantes durante el análisis



# Missing Values ...

---

- *Missing completely at random* (**MCAR**)
  - El faltante es independiente de los atributos
  - Completar basándose en el atributo
  - El análisis en conjunto puede ser insesgado
- *Missing at Random* (**MAR**)
  - El faltante se relaciona con otras variables
  - Completar valores basándose en otros valores
  - Casi siempre produce un sesgo en el análisis
- *Missing Not at Random* (**MNAR**)
  - El faltante se debe a mediciones no observadas
  - El faltante puede ser informativo o no-ignorable
- Imposible determinar la situación a partir de los datos

# Datos Duplicados

---

- El conjunto de datos puede incluir objetos que son duplicados o, casi duplicados unos de otros
  - Un problema serio cuando se fusionan datos provenientes de fuentes heterogéneas
- Ej:
  - La misma persona con múltiples direcciones de correo electrónico
- Limpieza de datos
  - Proceso de tratamiento para datos duplicados
- Cuando deben removerse los datos duplicados?

# Medidas de Similitud y de Disimilitud (diferencia)

---

## □ Medida de Similitud

- Medida Numérica de cuan parecidos son dos objetos entre si.
- Es mayor cuanto mas parecidos son dos objetos entre si.
- Usualmente comprendida en el rango  $[0,1]$

## □ Medida de Similitud

- Medida numérica de cuan diferentes son dos objetos de entre si
- Es mas pequeña cuando los objetos son mas parecidos
- Usualmente la disimilitud mínima se representa con 0
- El limite superior varia

□ **Proximidad** se refiere a similitud o disimilitud.

# Similitud / Disimilitud para Atributos Simples

---

Similitud y Disimilitud de dos objetos,  $x$  e  $y$ , con respecto a un único atributo simple.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d =  x - y  / (n - 1)$ (values mapped to integers 0 to $n-1$ , where $n$ is the number of values)	$s = 1 - d$
Interval or Ratio	$d =  x - y $	$s = -d, s = \frac{1}{1+d}, s = e^{-d},$ $s = 1 - \frac{d - \min\_d}{\max\_d - \min\_d}$

# Distancia Euclídea

---

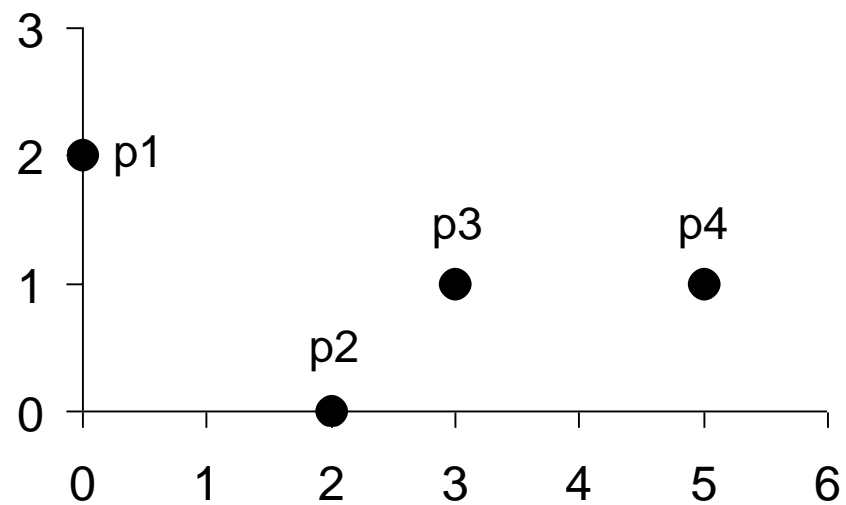
## □ Distancia Euclídea

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

donde  $n$  es la cantidad de dimensiones (atributos) y  $x_k$  e  $y_k$  son, respectivamente, el  $k^{esimo}$  atributo (componente) de los objetos  $\mathbf{x}$  e  $\mathbf{y}$ .

□ La estandarización es necesaria, si las escalas difieren.

# Distancia Euclidea



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Matriz de Distancias

# Distancia Minkowski

---

- Se trata de una generalización de la distancia Euclídea

$$d(\mathbf{x}, \mathbf{y}) = \left( \sum_{k=1}^n |x_k - y_k|^r \right)^{1/r}$$

donde  $r$  es un parámetro,  $n$  es la cantidad de dimensiones (atributos) y  $x_k$  e  $y_k$  son, respectivamente, el  $k^{esimo}$  atributo (componente) de los objetos  $x$  e  $y$ .

# Distancia Minkowski: Ejemplos

---

- $r = 1$ . Distancia Manhattan (City block, taxicab, etc.) norma  $L_1$ .
  - Un ejemplo usual es la distancia de Hamming, que es la cantidad de bits que son diferentes entre dos vectores binarios.
- $r = 2$ . Distancia euclídea.
- $r \rightarrow \infty$ . Distancia suprema o “*supremum*” (norma  $L_{\max}$ , norma  $L_{\infty}$ ).
  - Esta es la máxima diferencia entre cualquier componente de los vectores
- Cuidado en no confundir  $r$  con  $n$ .



# Distancia Minkowski

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

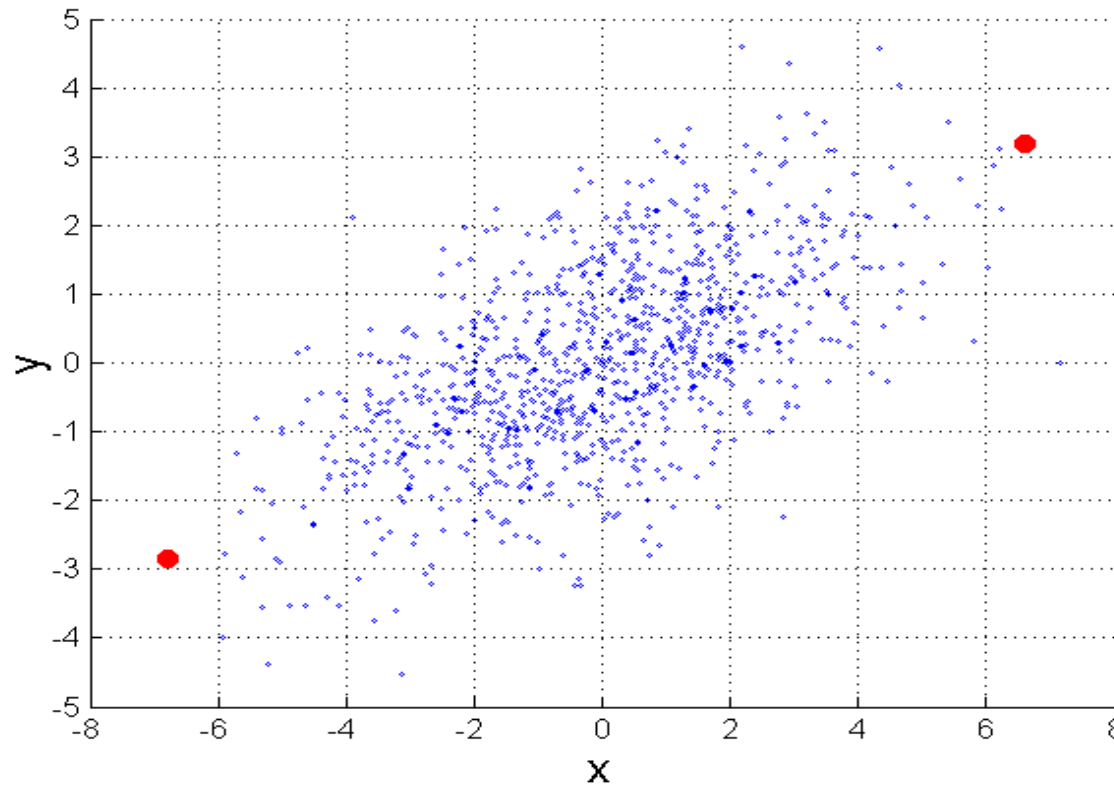
$L_{\infty}$	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

## Matriz de Distancia

# Distancia Mahalanobis

---

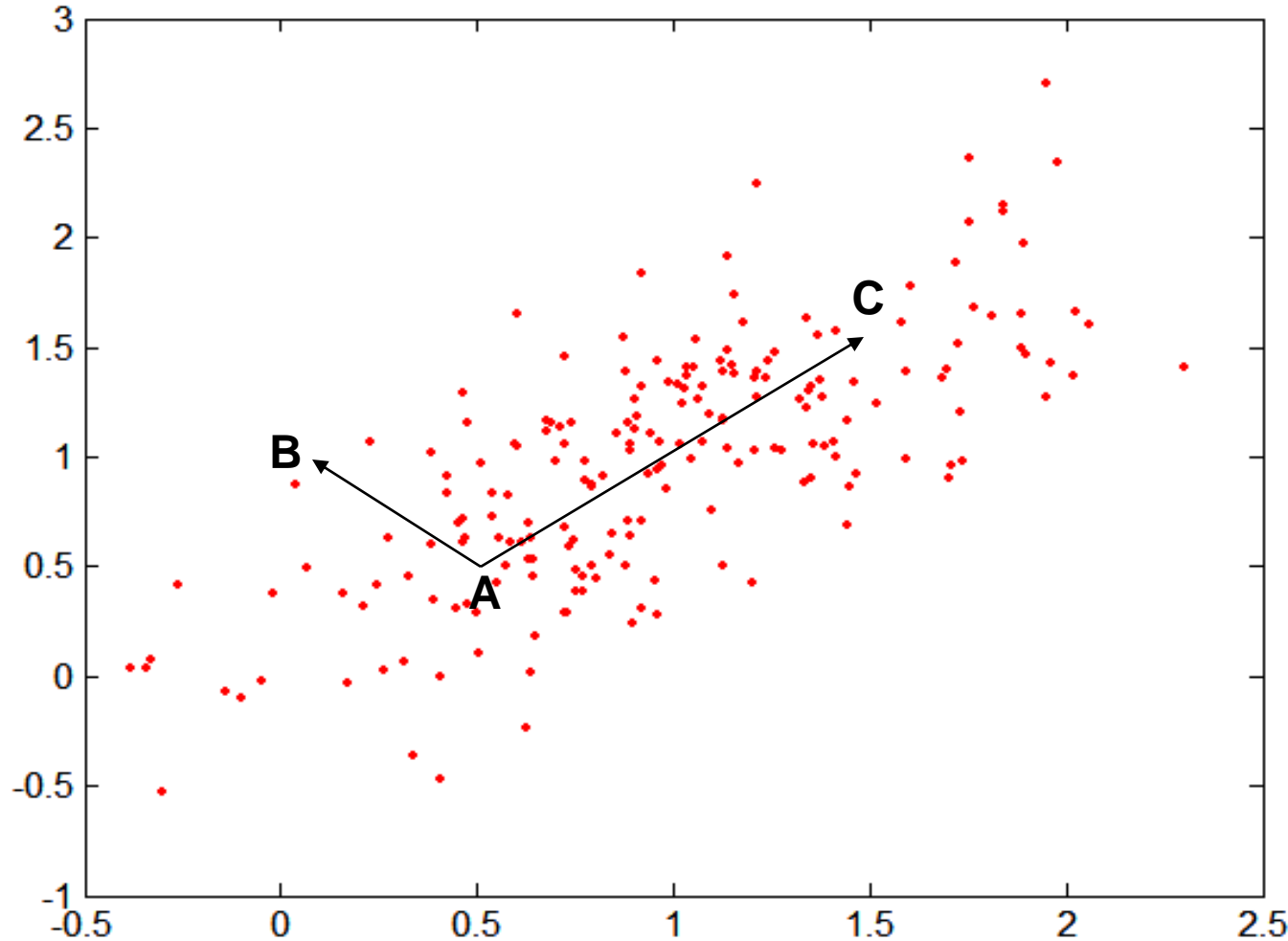
$$\text{mahalanobis}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y})}$$



$\Sigma$  es la matriz de covarianza

Para los puntos rojos, la distancia euclídea es 14.7, mientras que Mahalanobis es 6.

# Distance Mahalanobis



**Matriz de Covarianza**

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

**A: (0.5, 0.5)**

**B: (0, 1)**

**C: (1.5, 1.5)**

**Mahal(A,B) = 5**

**Mahal(A,C) = 4**

# Propiedades comunes de una distancia

---

- Las distancias tienen algunas propiedades conocidas.
  1.  $d(\mathbf{x}, \mathbf{y}) \geq 0$  para todo  $\mathbf{x}$  e  $\mathbf{y}$ . Además  $d(\mathbf{x}, \mathbf{y}) = 0$  sii  $\mathbf{x} = \mathbf{y}$ . (Definida Positiva)
  2.  $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$  para todo  $\mathbf{x}$  e  $\mathbf{y}$ . (Simetría)
  3.  $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$  para todos los puntos  $\mathbf{x}$ ,  $\mathbf{y}$ , y  $\mathbf{z}$ . (Desigualdad triangular)

donde  $d(\mathbf{x}, \mathbf{y})$  es la distancia (disimilitud) entre los puntos (objetos),  $\mathbf{x}$  e  $\mathbf{y}$ .

- Una distancia que satisface estas propiedades se denomina una **métrica**

# Propiedades comunes de una Similitud

---

□ Las similitudes también tienen propiedades conocidas.

1.  $s(\mathbf{x}, \mathbf{y}) = 1$  (o similitud máxima) sii  $\mathbf{x} = \mathbf{y}$ .

2.  $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$  para todo  $\mathbf{x}$  e  $\mathbf{y}$ . (Simetría)

donde  $s(\mathbf{x}, \mathbf{y})$  es la similitud entre los puntos (objetos),  $\mathbf{x}$  e  $\mathbf{y}$ .

# Similitud entre vectores binarios

---

- Situación común en la que los objetos,  $p$  y  $q$ , tienen solo atributos binarios

- La similitud se calcula empleando

$f_{01}$  = la cantidad de atributos donde  $p = 0$  y  $q = 1$

$f_{10}$  = la cantidad de atributos donde  $p = 1$  y  $q = 0$

$f_{00}$  = la cantidad de atributos donde  $p = 0$  y  $q = 0$

$f_{11}$  = la cantidad de atributos donde  $p = 1$  y  $q = 1$

- Coeficiente de Matcheo Simple (*Simple Matching*) y de Jaccard

SMC = cantidad de coincidencias / cantidad de atributos

$$= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$$

J = cantidad de coincidencias de tipo 11 / cantidad de atributos no nulos

$$= (f_{11}) / (f_{01} + f_{10} + f_{11})$$

# Ejemplo: SMC versus Jaccard

---

$$\mathbf{x} = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$$

$$\mathbf{y} = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$$

$$f_{01} = 2 \quad (\text{la cantidad de atributos donde } p = 0 \text{ y } q = 1)$$

$$f_{10} = 1 \quad (\text{la cantidad de atributos donde } p = 1 \text{ y } q = 0)$$

$$f_{00} = 7 \quad (\text{la cantidad de atributos donde } p = 0 \text{ y } q = 0)$$

$$f_{11} = 0 \quad (\text{la cantidad de atributos donde } p = 1 \text{ y } q = 1)$$

$$\begin{aligned} \text{SMC} &= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00}) \\ &= (0+7) / (2+1+0+7) = \mathbf{0.7} \end{aligned}$$

$$\mathbf{J} = (f_{11}) / (f_{01} + f_{10} + f_{11}) = 0 / (2 + 1 + 0) = \mathbf{0}$$

# Similitud Coseno

---

□ Si  $\mathbf{d}_1$  y  $\mathbf{d}_2$  son dos vectores (por ejemplo vectores de documentos), entonces

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \|\mathbf{d}_2\| ,$$

donde  $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$  indica el producto interno de los vectores,  $\mathbf{d}_1$  y  $\mathbf{d}_2$ , mientras  $\|\mathbf{d}\|$  es la longitud del vector  $\mathbf{d}$ . (recordemos que  $\|\mathbf{d}\| = \sqrt{\langle \mathbf{d}, \mathbf{d} \rangle} = \sqrt{(d_1*d_1 + d_2*d_2 + \dots d_n*d_n)}$ )

□ Ej.:

$$\mathbf{d}_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

$$\mathbf{d}_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$

$$\langle \mathbf{d}_1, \mathbf{d}_2 \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$\|\mathbf{d}_1\| = (3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$\|\mathbf{d}_2\| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0.5} = (6)^{0.5} = 2.449$$

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0.3150$$



# Extended Jaccard Coefficient (Tanimoto)

---

- Es una variación de Jaccard para atributos continuos o de conteo
  - Se reduce a Jaccard para atributos binarios

$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

# La correlación mide la relación lineal entre objetos

---

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x s_y}, \quad (2.11)$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) \quad (2.12)$$

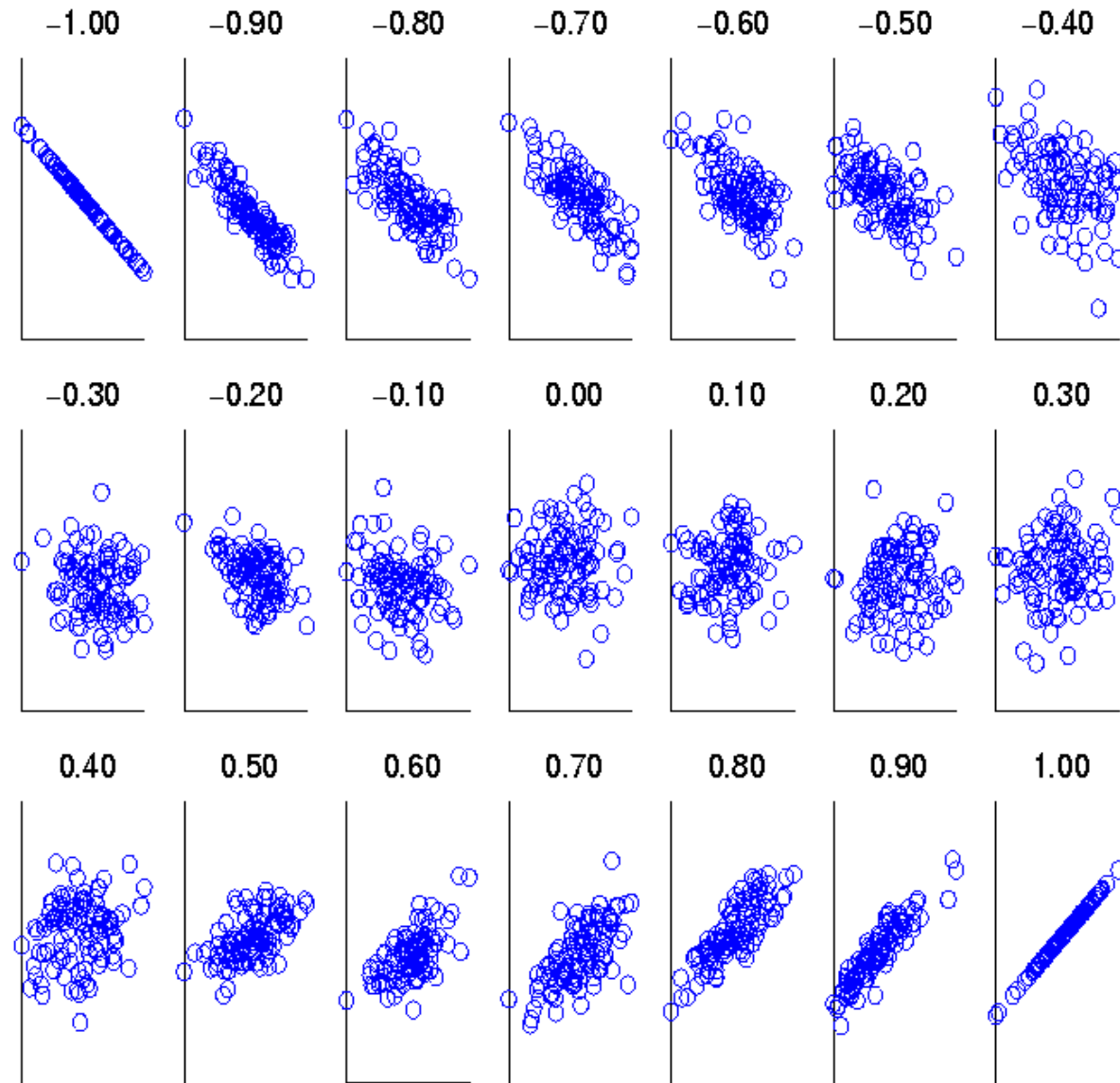
$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k \text{ is the mean of } \mathbf{x}$$

$$\bar{y} = \frac{1}{n} \sum_{k=1}^n y_k \text{ is the mean of } \mathbf{y}$$

# Evaluación Visual de la correlación para dos variables



**Gráficos de dispersión  
mostrando similitud de -1 a 1**

# Problemas con la correlación

---

□  $\mathbf{x} = (-3, -2, -1, 0, 1, 2, 3)$

□  $\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$

$$y_i = x_i^2$$

□  $\text{mean}(\mathbf{x}) = 0, \quad \text{mean}(\mathbf{y}) = 4$

□  $\text{std}(\mathbf{x}) = 2.16, \quad \text{std}(\mathbf{y}) = 3.74$

□  $\text{corr} = (-3)(5) + (-2)(0) + (-1)(-3) + (0)(-4) + (1)(-3) + (2)(0) + 3(5) / (6 * 2.16 * 3.74)$   
 $= 0$

# Comparación de las Medidas de Proximidad

---

- Dominio de aplicación
  - Las medidas de similitud tienden a ser específicas para el tipo de atributo y de dato
  - Para registros, imágenes, grafos, secuencias, etc. tienden a ser diferentes
- Aun así, es usual considerar las propiedades que uno pretende que una medida de proximidad tenga
  - Simetría es una propiedad usual
  - Tolerancia al ruido y a los valores extremos es otra
  - Habilidad de encontrar mas tipos de patrones?
  - Muchas otras opciones...
- La medida debe ser aplicable a los datos en cuestión y producir resultados congruentes con el dominio de conocimiento

# Medidas Basadas en la Información

---

- Teoría de la información como disciplina bien desarrollada y fundamental en múltiples aplicaciones.
- Algunas medidas de similitud están basadas en teoría de la información
  - Información mutua (en varias versiones)
  - Coeficiente de Información Máxima (*Maximal Information Coefficient* (**MIC**)) y medidas relacionadas
  - Es general y puede emplearse también para relaciones no lineales
  - Pueden ser complicadas y computacionalmente costosas

# Información y Probabilidad

---

- La información se relaciona con los resultados posibles de un evento
  - Transmisión de un mensaje, tiro de una moneda, medición de una parte de los datos



- Cuanto mas seguro un resultado, menor información contiene y viceversa
  - Por ejemplo, si una moneda posee dos caras, entonces el resultado de obtener una cara no provee ninguna información
  - En forma mas cuantitativa, la información esta relacionada con la probabilidad de un resultado
    - ◆ Cuanta menor la probabilidad de un resultado, mayor es la información que provee y viceversa
  - Entropía es la medida mas usualmente empleada

# Entropía

---

## □ Para

- una variable (evento),  $X$ ,
- con  $n$  valores posibles (resultados),  $x_1, x_2, \dots, x_n$
- cada resultado con una probabilidad dada por  $p_1, p_2, \dots, p_n$
- La entropía de  $X$ ,  $H(X)$ , está dada por

$$H(X) = - \sum_{i=1}^n p_i \log_2 p_i$$

## □ La entropía tiene un valor entre 0 y $\log_2 n$ y se mide en bits

- Por lo tanto, la entropía es una medida de cuantos bits se necesitan para representar en promedio una observación de  $X$



# Ejemplos de entropía

---

- Para una moneda con una probabilidad
  - $p$  de cara y probabilidad  $q = 1 - p$  de seca

$$H = -p \log_2 p - q \log_2 q$$

- Para  $p = 0.5$ ,  $q = 0.5$  (moneda justa)  $H = 1$
  - Para  $p = 1$  o  $q = 1$ ,  $H = 0$
- Cual es la entropía de un dado justo de cuatro caras?

# Entropía para datos de ejemplo

---

Hair Color	Count	$p$	$-p\log_2 p$
Black	75	0.75	0.3113
Brown	15	0.15	0.4105
Blond	5	0.05	0.2161
Red	0	0.00	0
Other	5	0.05	0.2161
Total	100	1.0	1.1540

La entropía máxima es  $\log_2 5 = 2.3219$

# Entropía para datos muestrales

---

## □ Asumamos

- una cantidad de observaciones ( $m$ ) de algún atributo,  $X$ , e.g., el color de cabello de los alumnos en una clase,
- donde hay  $n$  valores diferentes posibles
- y el numero de observaciones en la  $i^{esima}$  categoría es  $m_i$
- entonces, para este ejemplo

$$H(X) = - \sum_{i=1}^n \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

- Para datos continuos, los cálculos son mas complejos.

# Información Mutua

---

- La información que una variable provee sobre otra

Formalmente,  $I(X, Y) = H(X) + H(Y) - H(X, Y)$ , donde

$H(X, Y)$  es la entropía conjunta de  $X$  e  $Y$ ,

$$H(X, Y) = - \sum_i \sum_j p_{ij} \log_2 p_{ij}$$

donde  $p_{ij}$  es la probabilidad que el  $i^{\text{ésimo}}$  valor de  $X$  y el  $j^{\text{ésimo}}$  valor de  $Y$  ocurran juntos

- Para variables discretas, es fácil de calcular
- La información mutua máxima para una variable discreta es

$\log_2(\min(n_X, n_Y))$ , donde  $n_X$  ( $n_Y$ ) es el número de valores de  $X$  ( $Y$ )

# Ejemplo de Información Mutua

Student Status	Count	$p$	$-p\log_2 p$
Undergrad	45	0.45	0.5184
Grad	55	0.55	0.4744
Total	100	1.00	0.9928

Grade	Count	$p$	$-p\log_2 p$
A	35	0.35	0.5301
B	50	0.50	0.5000
C	15	0.15	0.4105
Total	100	1.00	1.4406

Student Status	Grade	Count	$p$	$-p\log_2 p$
Undergrad	A	5	0.05	0.2161
Undergrad	B	30	0.30	0.5211
Undergrad	C	10	0.10	0.3322
Grad	A	30	0.30	0.5211
Grad	B	20	0.20	0.4644
Grad	C	5	0.05	0.2161
Total		100	1.00	2.2710

Información Mutua de *Student Status* y *Grade* =  $0.9928 + 1.4406 - 2.2710 = 0.1624$

# Coeficiente de Información Maximal (*Maximal Information Coefficient*)

---

- Reshef, David N., Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. "Detecting novel associations in large data sets." *science* 334, no. 6062 (2011): 1518-1524.
- Aplica la información mutua a dos variables continuas
- Considera el posible agrupamiento de las variables en categorías discretas
  - $n_X \times n_Y \leq N^{0.6}$  donde
    - ◆  $n_X$  es la cantidad de valores de  $X$
    - ◆  $n_Y$  es la cantidad de valores de  $Y$
    - ◆  $N$  es la cantidad de muestras (observaciones, objetos)
- Calcula la información mutua
  - Normalizada por  $\log_2(\min(n_X, n_Y))$
- Toma el mayor valor

# Aproximación General a la Combinación de Similitudes

---

- Cuando los atributos son de diferentes tipos pero se necesita una medida de similitud general.

1: Para el  $k^{\text{esimo}}$  atributo, calcular una similitud,  $s_k(\mathbf{x}, \mathbf{y})$ , con valores en el rango  $[0, 1]$ .

2: Definir una variable indicadora,  $\delta_k$ , para el  $k^{\text{esimo}}$  atributo de la siguiente forma:

$\delta_k = 0$  si el  $k^{\text{esimo}}$  atributo es un atributo asimétrico y ambos objetos tienen un valor de 0, o si uno de los objetos tiene un valor faltante para el  $k^{\text{esimo}}$  atributo

$\delta_k = 1$  sino

3. Calcular

$$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \delta_k}$$

# Ponderadores para combinar similitudes

---

- Es posible que no queramos tratar a todos los atributos de la misma forma.

- Utilizamos entonces pesos no-negativos  $\omega_k$

- $similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n \omega_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \omega_k \delta_k}$

- También se puede definir una forma ponderada de la distancia

$$d(\mathbf{x}, \mathbf{y}) = \left( \sum_{k=1}^n w_k |x_k - y_k|^r \right)^{1/r}$$



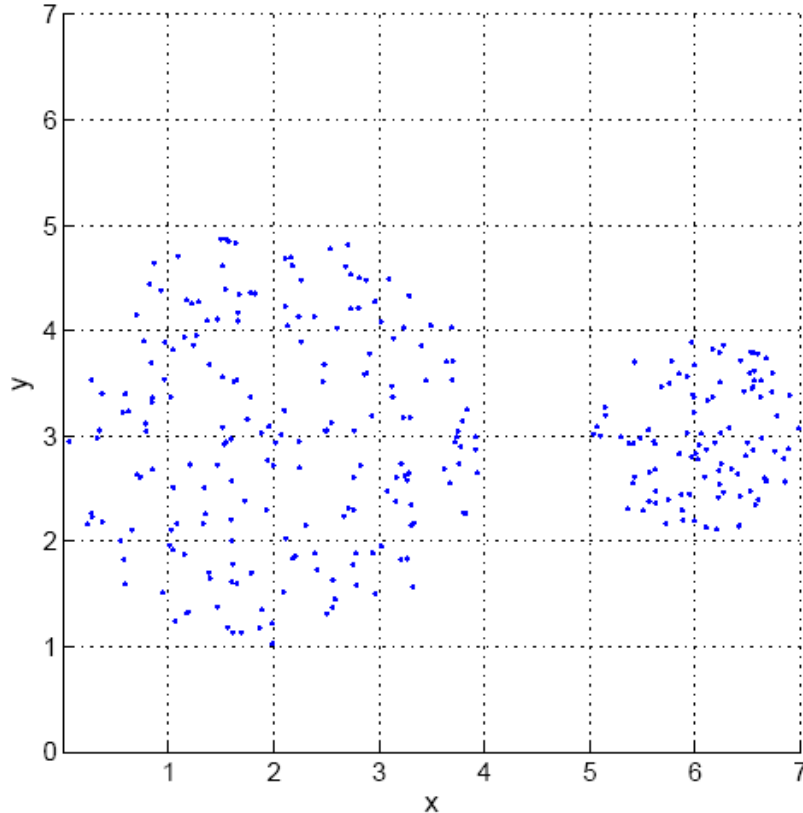
# Densidad

---

- Mide el grado en el que los objetos están mas cercanos unos a otros en un área especificada
- La noción de densidad esta fuertemente relacionada con la de proximidad
- El concepto de densidad se emplea típicamente para clustering y para detección de anomalías
- Ejemplos:
  - Densidad euclidiana
    - ◆ DE = cantidad de puntos por unidad de volumen
  - Densidad de Probabilidad
    - ◆ Requiere estimar la forma de la distribución de los datos
  - Densidad basada en el grafo
    - ◆ Conectividad

# Densidad Euclidiana: Aproximación basada en una grilla

- La forma mas simple es dividir la región en un conjunto de celdas de igual volumen y definir la densidad como la cantidad de puntos que contiene cada celda...



**Grid-based density.**

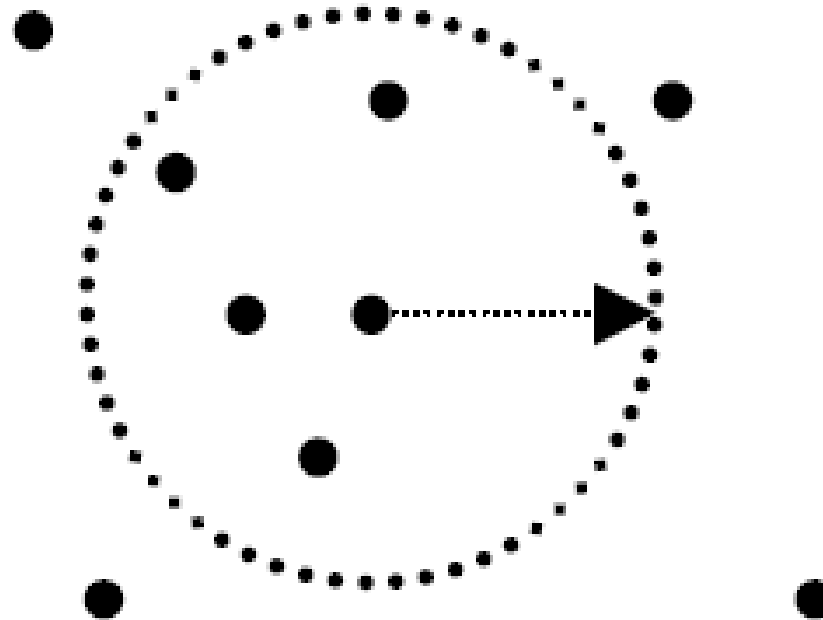
0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

**Counts for each cell.**

# Densidad Euclidiana: *Center-Based*

---

- En este caso se calcula la cantidad de puntos dentro un dado radio a partir del punto



**Center-based density**

# Preprocesamiento de Datos

---

- Agregación
- Muestreo
- Reducción de la Dimensionalidad
- Selección de Características
- Extracción de Características
- Discretización y Binarización
- Transformación de Atributos

# Agregación

---

□ Combinar dos o mas atributos (u objetos) en un único atributo (u objeto)

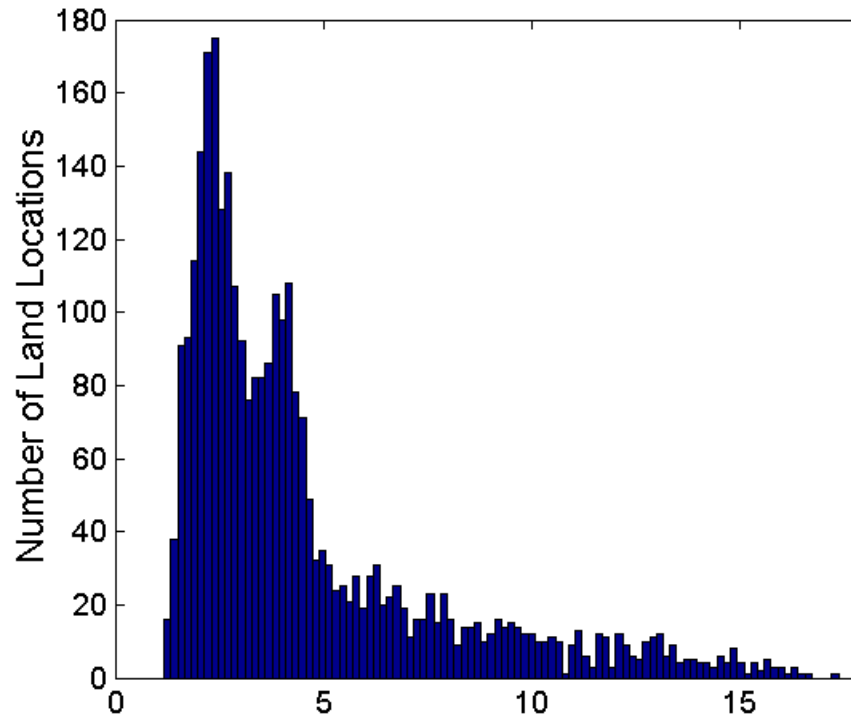
□ **Propósito**

- Reducción de datos
  - ◆ Reducir la cantidad de atributos u objetos
- Cambio de escala
  - ◆ Ciudades agregadas en regiones, estados, países, etc.
  - ◆ Días agregados en semanas, meses o años
- Hacer mas “estables” los datos
  - ◆ Los datos agregados tienden a tener menor variabilidad

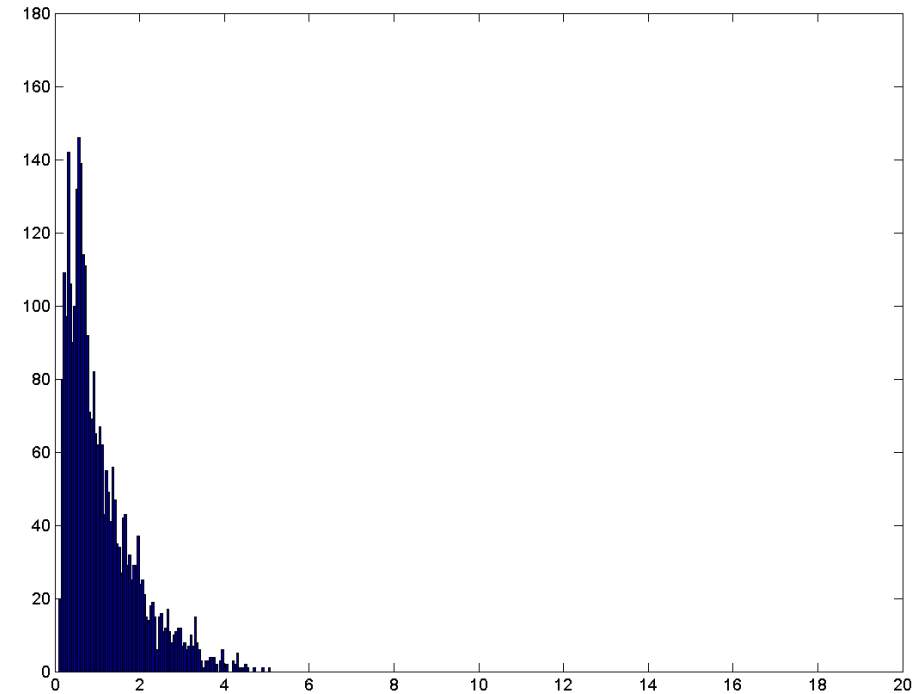
# Ejemplo: Precipitación en Australia ...

---

## Variación de las Precipitaciones en Australia



**Standard Deviation of Average  
Monthly Precipitation**



**Standard Deviation of  
Average Yearly Precipitation**

# Muestreo

---

- El muestreo es la principal técnica empleada para la reducción de datos.
  - Se la suele emplear tanto para una investigación preliminar de los datos como para el análisis final.
- Usualmente se muestrea debido a que **obtener** el conjunto completo de datos de interés es excesivamente caro o muy engorroso en tiempo.
- El muestreo se utiliza típicamente en minería de datos porque **procesar** el conjunto de datos completo es muy caro o engorroso en tiempo.

# Muestreo ...

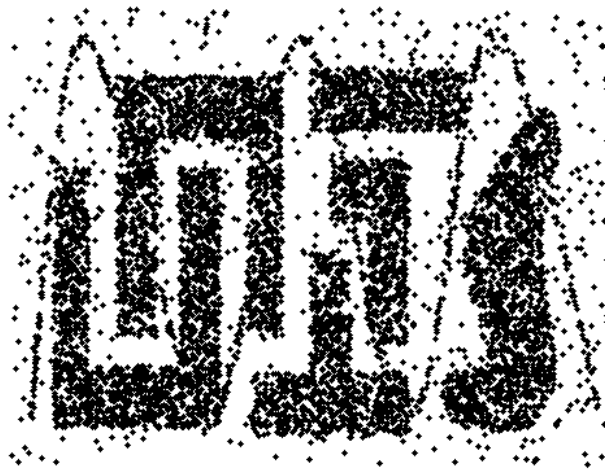
---

- El principio clave para un muestreo efectivo es:
  - El empleo de la muestra funciona casi tan bien como utilizar todo el conjunto de datos si la muestra es **representativa**
  - Una muestra es **representativa** si tiene aproximadamente las mismas propiedades (de interés) que el conjunto original de datos.

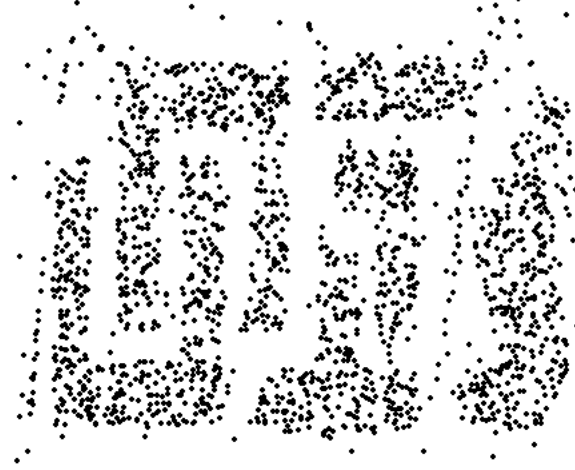


# Tamaño de la muestra

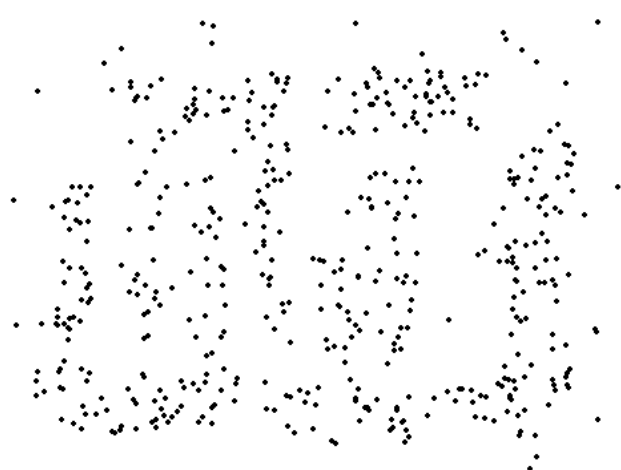
---



8000 puntos



2000 puntos



500 puntos

# Tipos de Muestreo

---

## □ Muestreo Aleatorio Simple

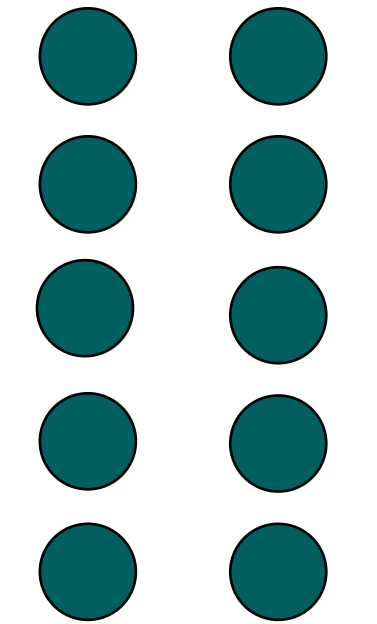
- Existe la misma probabilidad de elegir cualquier elemento en particular
- Muestreo sin reemplazo
  - ◆ En la medida que un elemento es seleccionado, se lo remueve de la población
- Muestreo con reemplazo
  - ◆ Los objetos **no** son removidos de la población en la medida que son seleccionados para la muestra.
  - ◆ El mismo objeto puede ser elegido mas de una vez para formar parte de la muestra

## □ Muestreo Estratificado

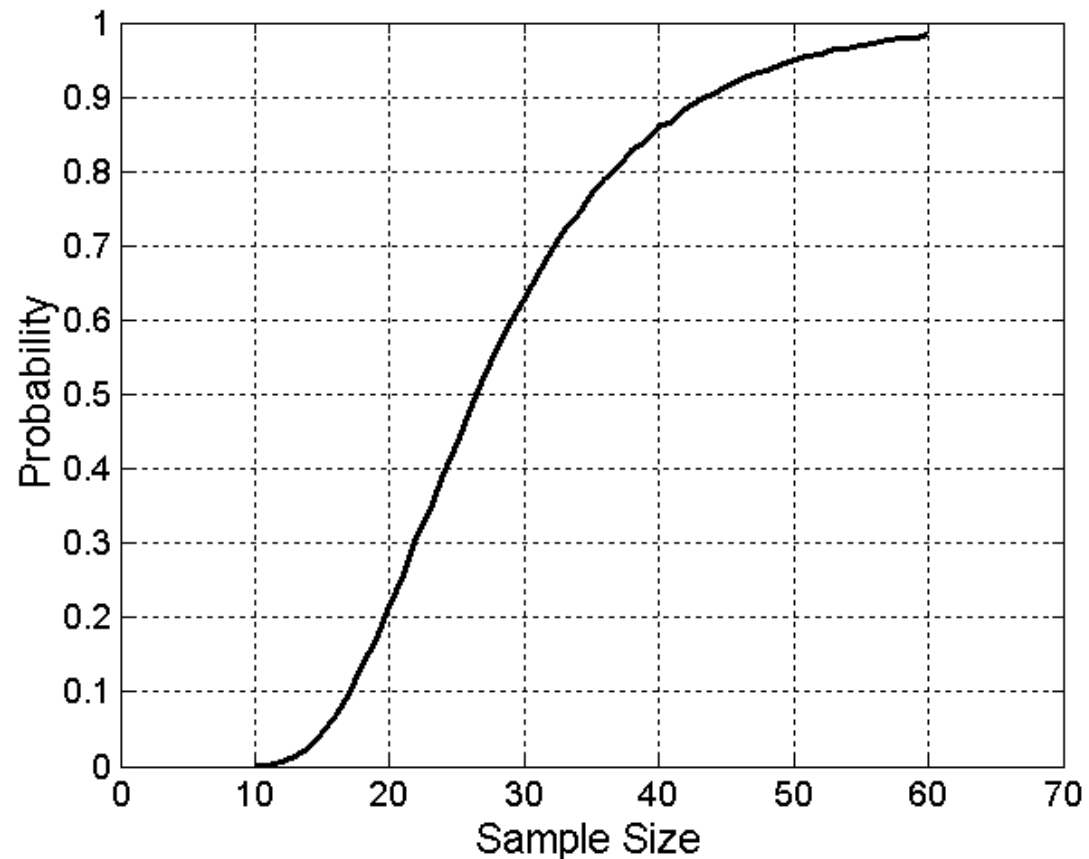
- El conjunto de datos se divide en varias particiones; luego se extraen muestras aleatorias dentro de cada partición.

# Tamaño de la muestra

- Cual es el tamaño de muestra necesario para obtener por lo menos un objeto de cada uno de 10 grupos de igual tamaño

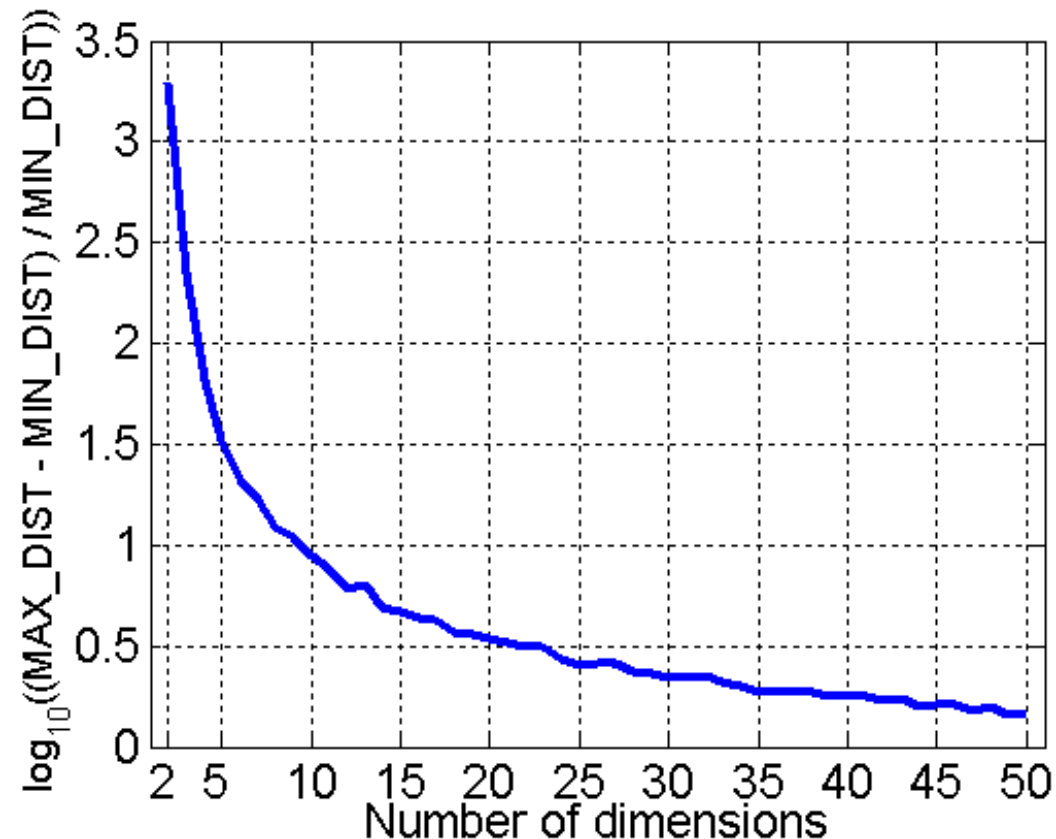


10 grupos de puntos



# La maldición de la dimensionalidad

- Cuando se incrementa la dimensionalidad los datos se convierten en cada vez mas escasos en el espacio que ocupan.
- Las definiciones de densidad y distancia entre puntos, que son criticas para conglomeración y detección de valores extremos comienzan a perder sentido



- Generar aleatoriamente 500 puntos
- Calcular la diferencia entre la distancia máxima y mínima entre pares de puntos

# Reducción de la Dimensionalidad

---

## □ Propósito:

- Evitar la maldición de la dimensionalidad
- Reducir el tiempo y memoria que requieren los algoritmos de minería de datos
- Permitir que los datos sean mas rápidamente visualizados
- Pueden ayudar a eliminar características irrelevantes o reducir el ruido

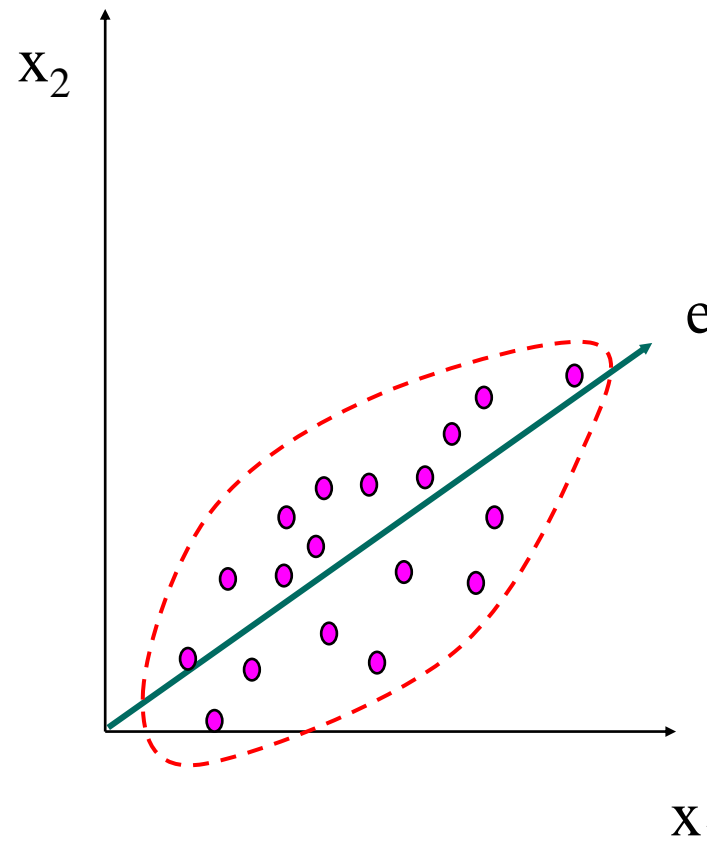
## □ Técnicas

- Análisis de Componentes Principales (ACP) *Principal Components Analysis (PCA)*
- Descomposición en Valores Singulares (DVS) *Singular Value Decomposition (SVD)*
- Otros: técnicas supervisadas y no-lineales

# Reducción de la Dimensionalidad: PCA

---

- El objetivo es encontrar una proyección que capture la mayor proporción de variación en los datos



# Reducción de la Dimensionalidad: PCA

---



# Selección de un Subconjunto de Características

---

- Otra forma de reducir la dimensionalidad de los datos
- Características redundantes
  - Duplican mucha o toda la información contenida en uno o mas atributos
  - Ejemplo: precio de compra de un producto y el monto de impuesto de venta pagado.
- Características Irrelevantes
  - No contienen ninguna información que es útil para la tarea de minería de datos en cuestión
  - Ejemplo: el ID del alumno es usualmente irrelevante para predecir el promedio de la carrera
- Muchas técnicas disponibles, especialmente desarrolladas en el contexto de la clasificación



# Creación de Características

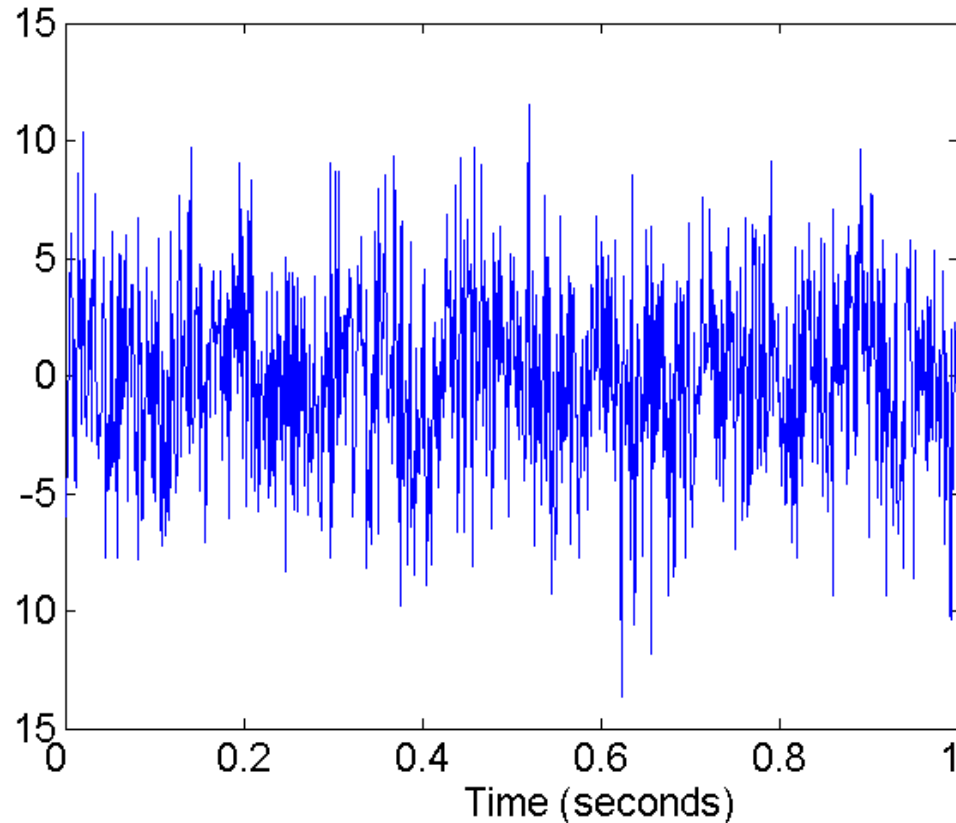
---

- Crear nuevos atributos que capturen la información importante en el conjunto de datos de forma mas eficiente que los atributos originales.
  
- Tres aproximaciones generales:
  - Extracción de características
    - ◆ Ejemplo: extracción de los bordes de los objetos en las imágenes
  - Construcción de características
    - ◆ Ejemplo: dividir la masa por el volumen para obtener la densidad
  - Mapear en un nuevo espacio
    - ◆ Ejemplo: Análisis de Fourier y de Wavelets

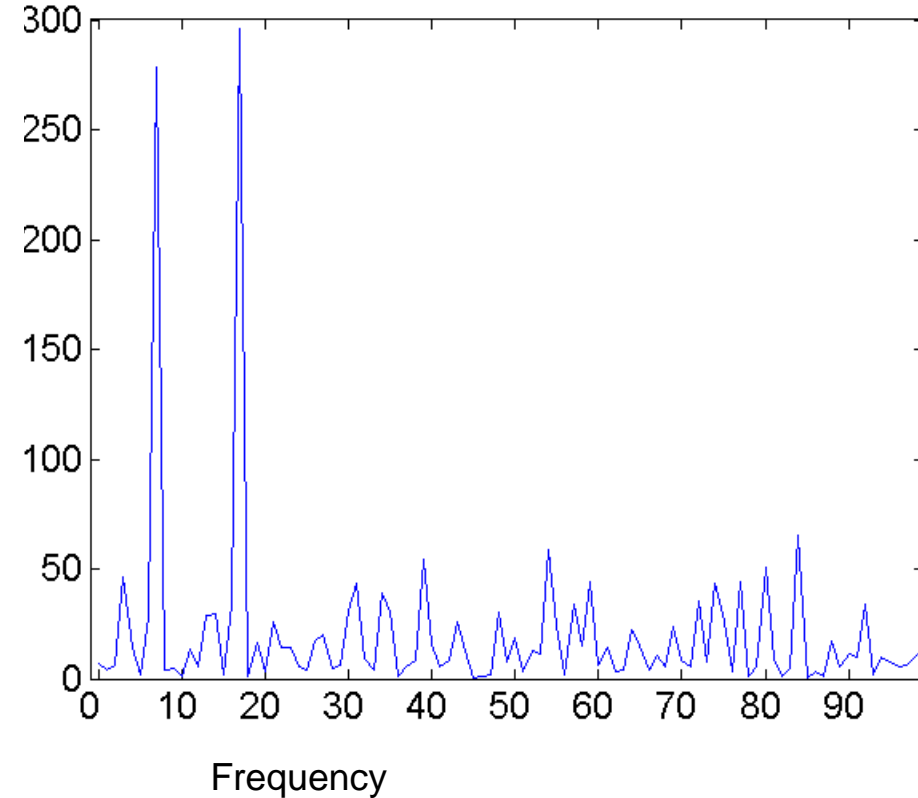
# Mapear Datos en un nuevo Espacio

---

## □ Transformaciones de Fourier y wavelets



**Two Sine Waves + Noise**



**Frequency**

# Discretización

---

- La Discretización es el proceso de convertir un atributo continuo en un atributo ordinal
  - Una cantidad potencialmente infinita de valores se mapean en una pequeña cantidad de categorías
  - La discretización se emplea usualmente en clasificación
  - Muchos de los algoritmos de clasificación trabajan mejor si tanto la variable independiente como las variables dependientes tienen poco valores posibles

# Iris Sample Data Set

---

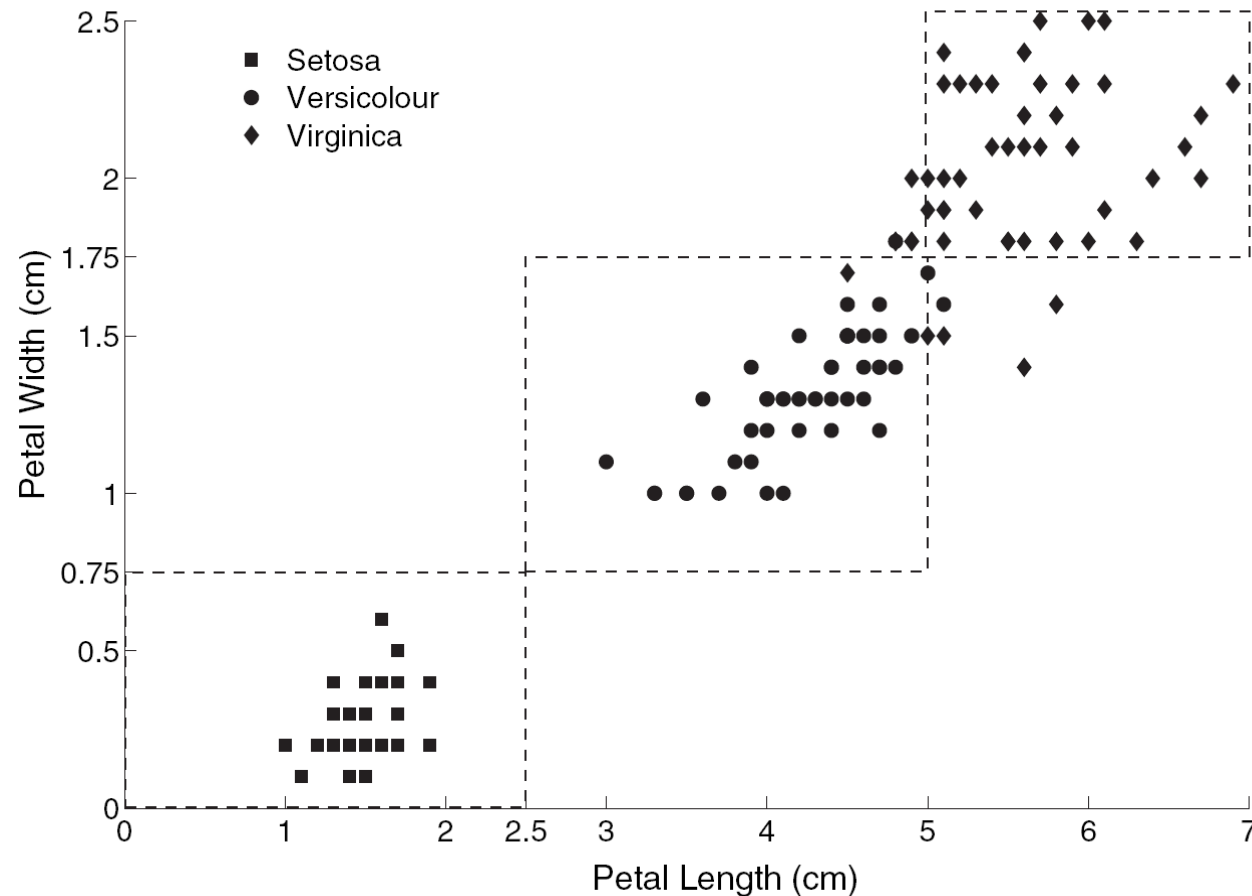
## □ Iris Plant data set.

- Se puede descargar del UCI Machine Learning Repository <http://www.ics.uci.edu/~mlearn/MLRepository.html>
- Provisto por Douglas Fisher
- Tres tipos de flores (clases):
  - ◆ Setosa
  - ◆ Versicolour
  - ◆ Virginica
- Cuatro atributos:
  - ◆ Sepal width and length
  - ◆ Petal width and length



Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

# Discretización: Ejemplo data set Iris



Petal width low or petal length low implies Setosa.

Petal width medium or petal length medium implies Versicolour.

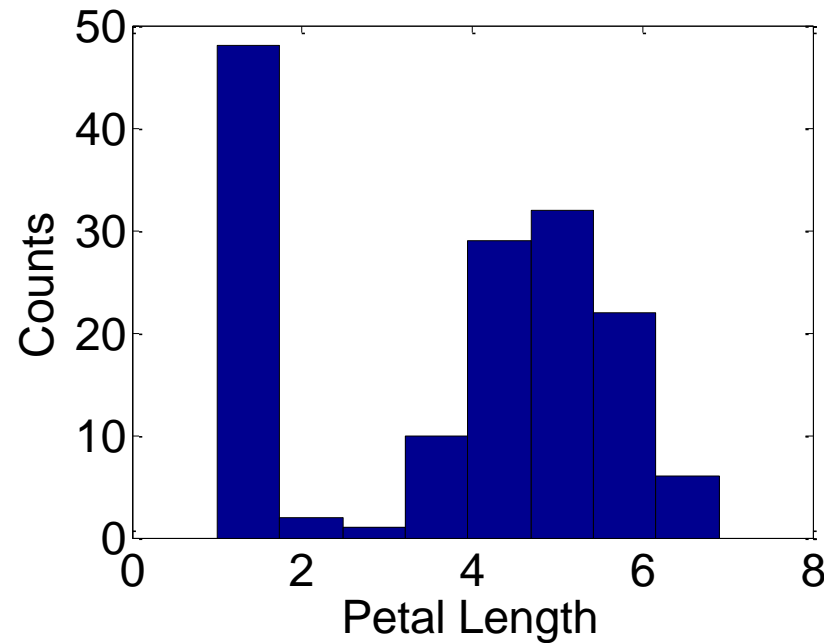
Petal width high or petal length high implies Virginica.

# Discretización: Iris data set ...

---

- Como podemos determinar cual es la mejor discretización?
  - Discretización NO-supervisada: encontrar cortes en los valores de los datos

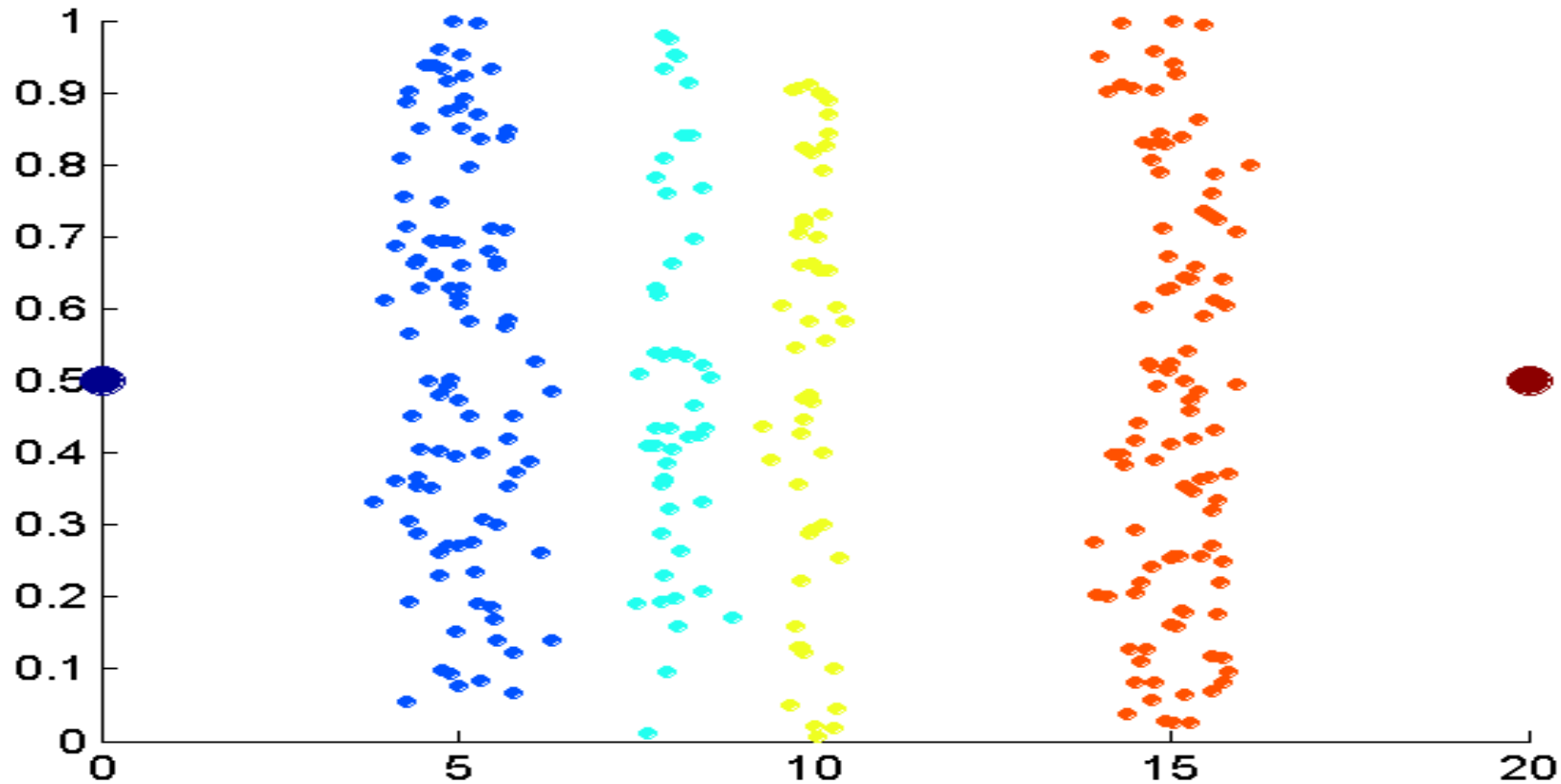
◆ Ej:  
Petal Length



- Discretización Supervisada: Utilizar la etiqueta de clase para encontrar los cortes

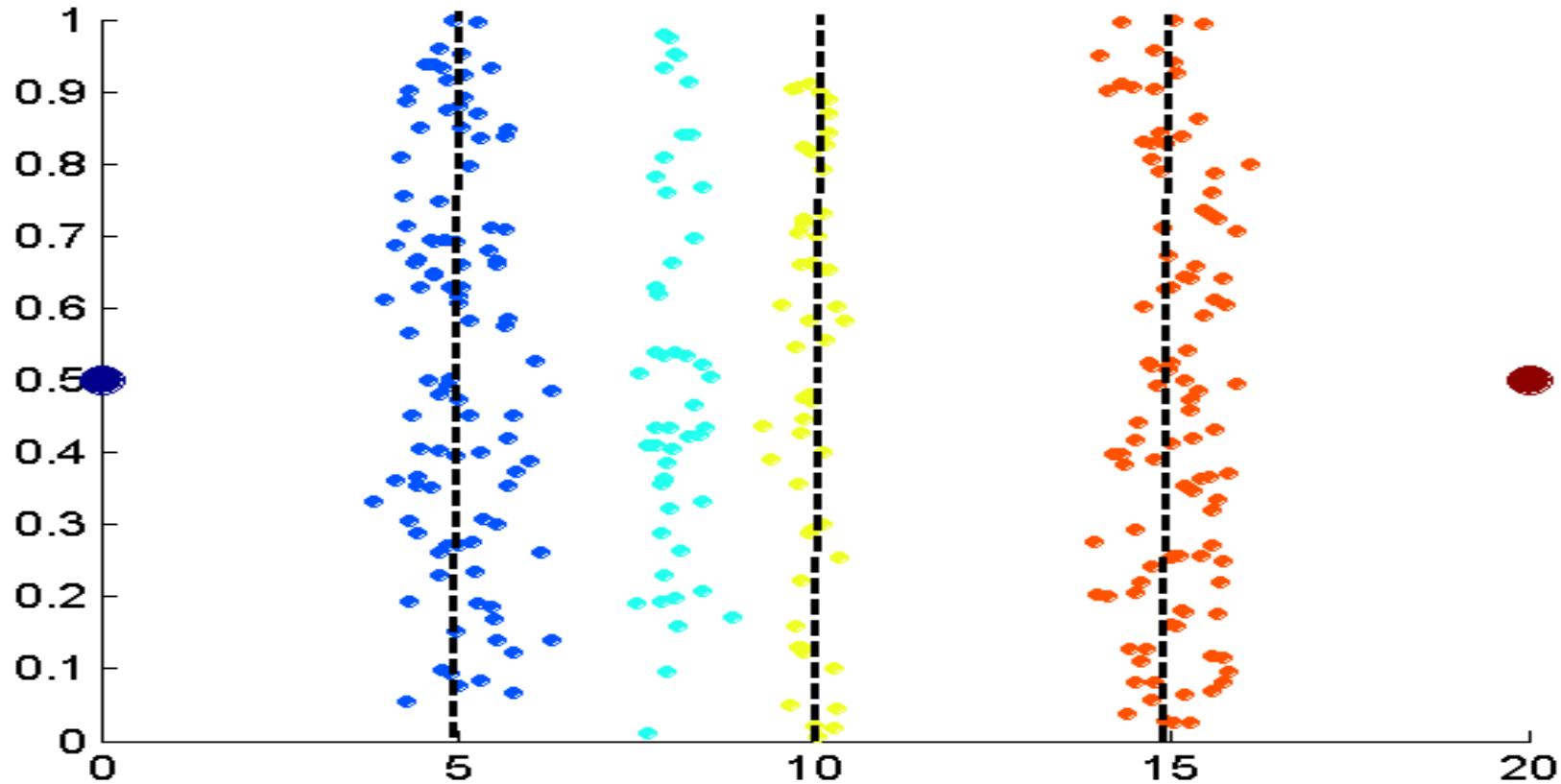
# Discretización sin utilizar etiquetas de clase

---



Los datos consisten en cuatro grupos de puntos y dos valores extremos. Se trata de datos unidimensionales, pero se agrego un componente aleatorio en y para reducir la superposición.

# Discretización sin utilizar etiquetas de clase

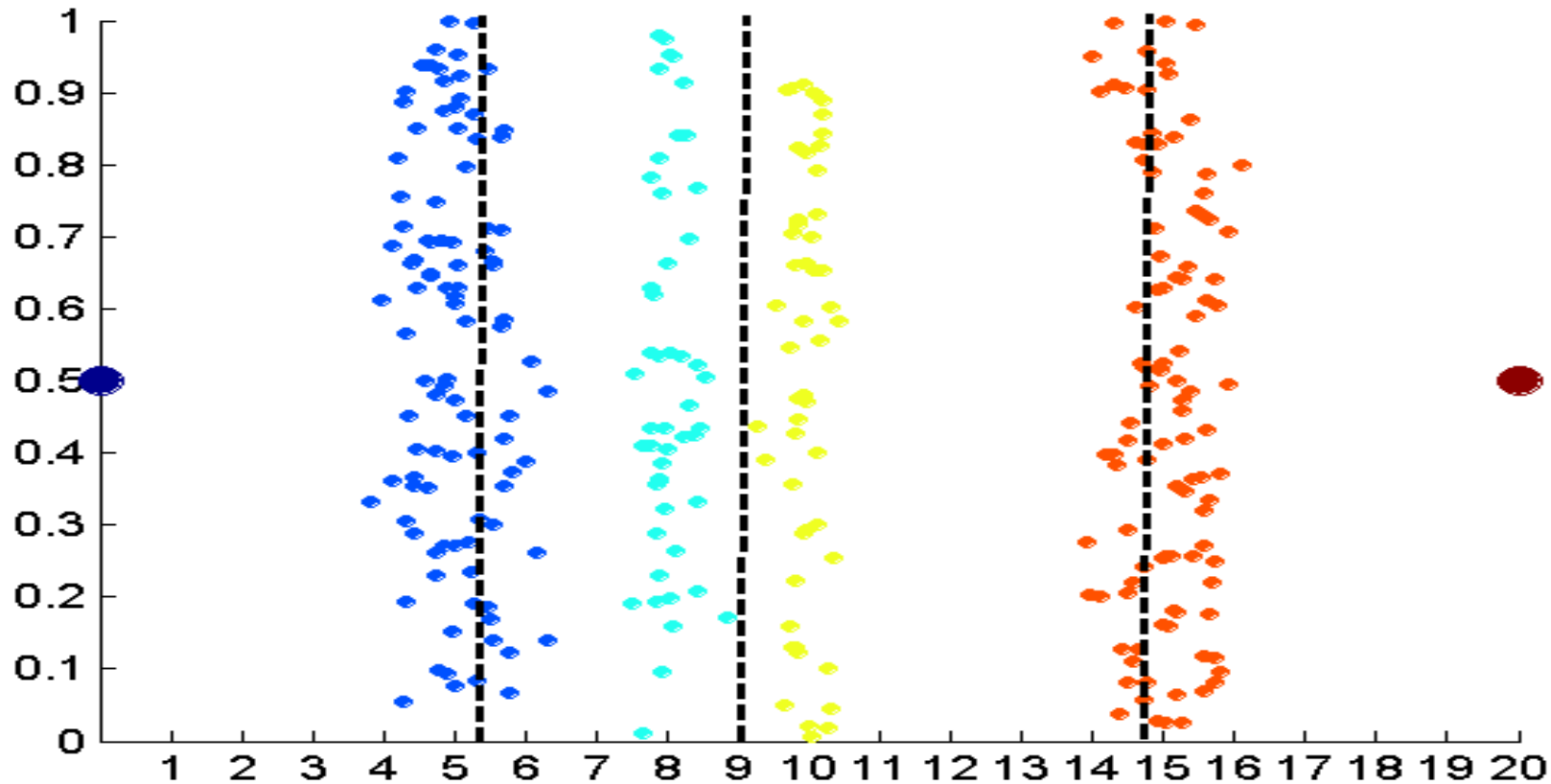


**Intervalos Iguales** para obtener cuatro valores



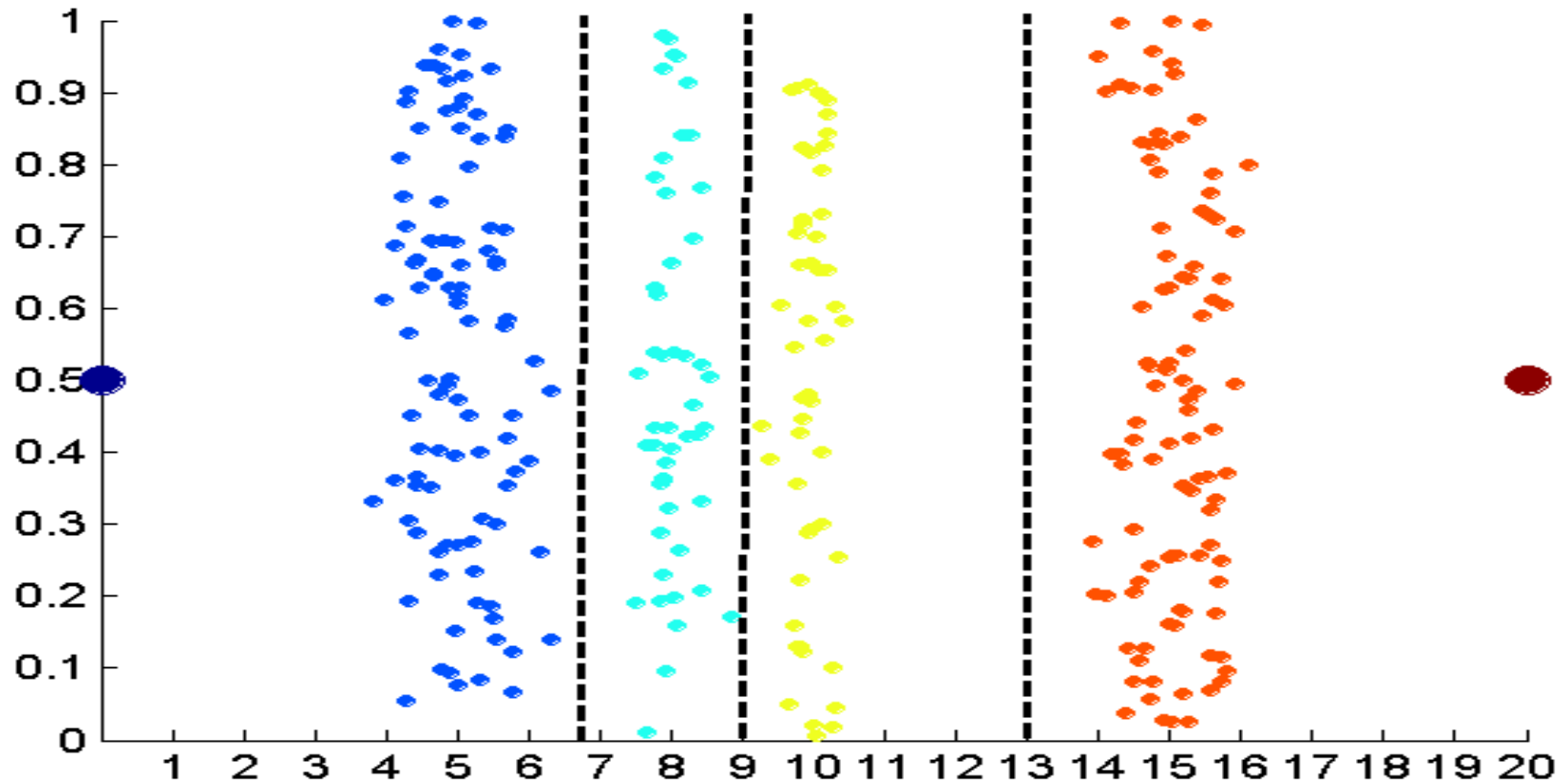
# Discretización sin utilizar intervalos de clase

---



**Igual Frecuencia** utilizada para obtener cuatro valores

# Discretización sin utilizar etiquetas de clase



K-means

# Binarizacion

---

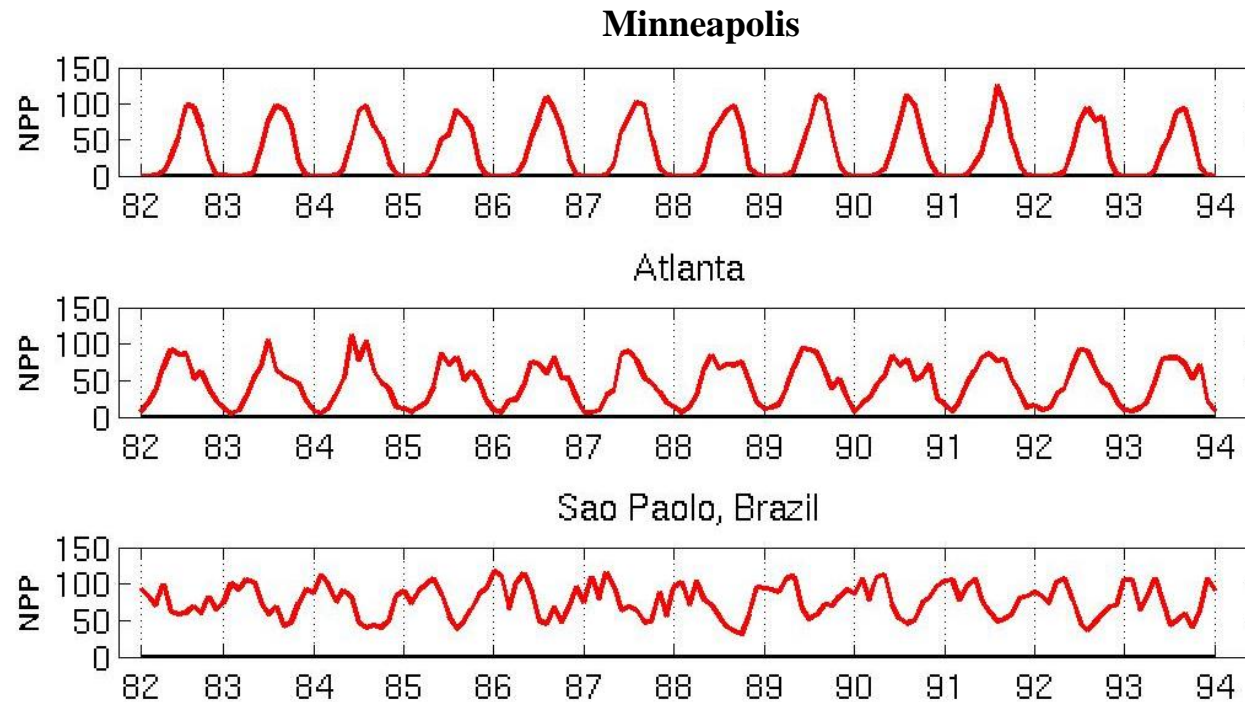
- La binarizacion mapea una variable continua o categórica en una o mas variables binarias
- Se emplea típicamente para el análisis de asociación
- Muchas veces se suele convertir primero de un atributo continuo en un atributo categórico y luego a un conjunto de atributos binarios
  - El análisis de asociación requiere atributos binarios asimetricos
  - Ejemplos: color de ojos y alturas medidos como {bajo, medio, alto}

# Transformación de Atributos

---

- Una **transformación de atributos** es una función que mapea el conjunto completo de valores de un dado atributo en un nuevo conjunto de valores de reemplazo de manera que cada uno de los valores antiguos pueda ser identificado con uno de los nuevos valores
  - Funciones simples:  $x^k$ ,  $\log(x)$ ,  $e^x$ ,  $|x|$
  - **Normalización**
    - ◆ Se refiere en términos generales a diversas técnicas que ajustan las diferencias entre los atributos en términos de la frecuencia de ocurrencia, la media, varianza, rango, etc.
    - ◆ Remueven componentes no deseados, o comunes, por ejemplo la estacionalidad
  - En estadística, la **estandarización** se refiere a restar la media y dividir por la desviación estándar.

# Ejemplo: Sample Time Series of Plant Growth

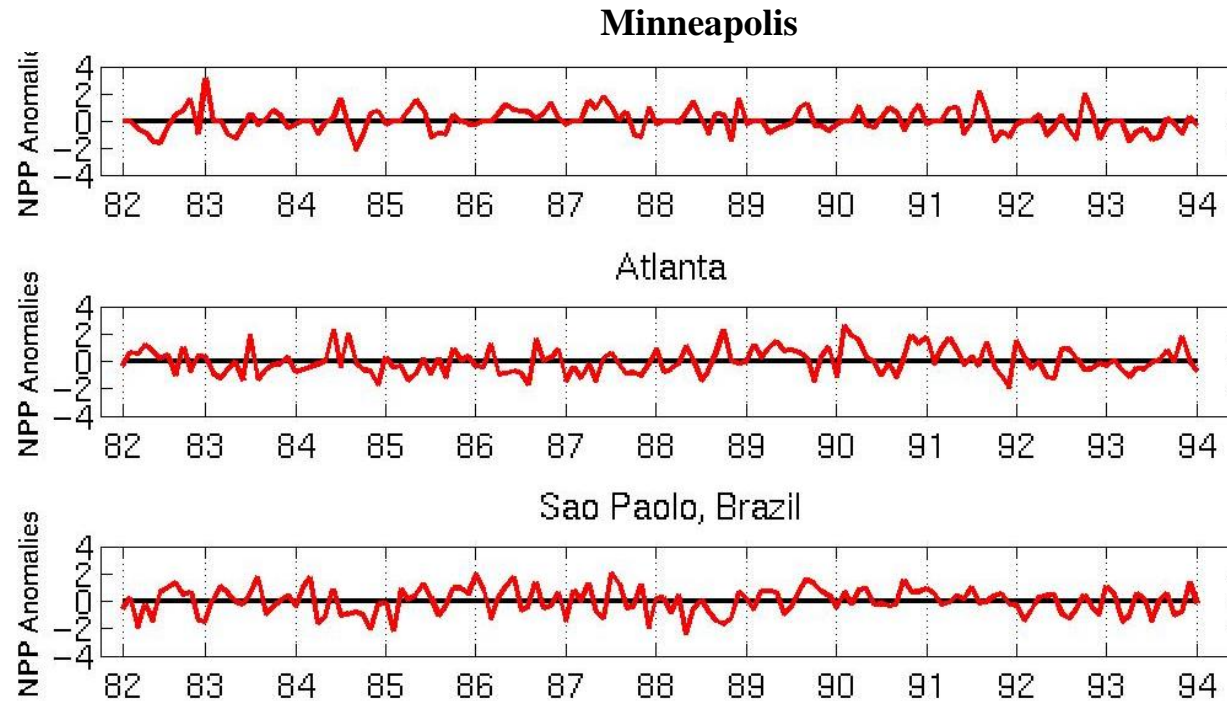


**Net Primary Production (NPP) is a measure of plant growth used by ecosystem scientists.**

## Correlations between time series

	Minneapolis	Atlanta	Sao Paulo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paulo	-0.7581	-0.5739	1.0000

# Seasonality Accounts for Much Correlation



Normalized using  
monthly Z Score:

Subtract off monthly  
mean and divide by  
monthly standard  
deviation

## Correlations between time series

	Minneapolis	Atlanta	Sao Paulo
Minneapolis	1.0000	0.0492	0.0906
Atlanta	0.0492	1.0000	-0.0154
Sao Paulo	0.0906	-0.0154	1.0000

