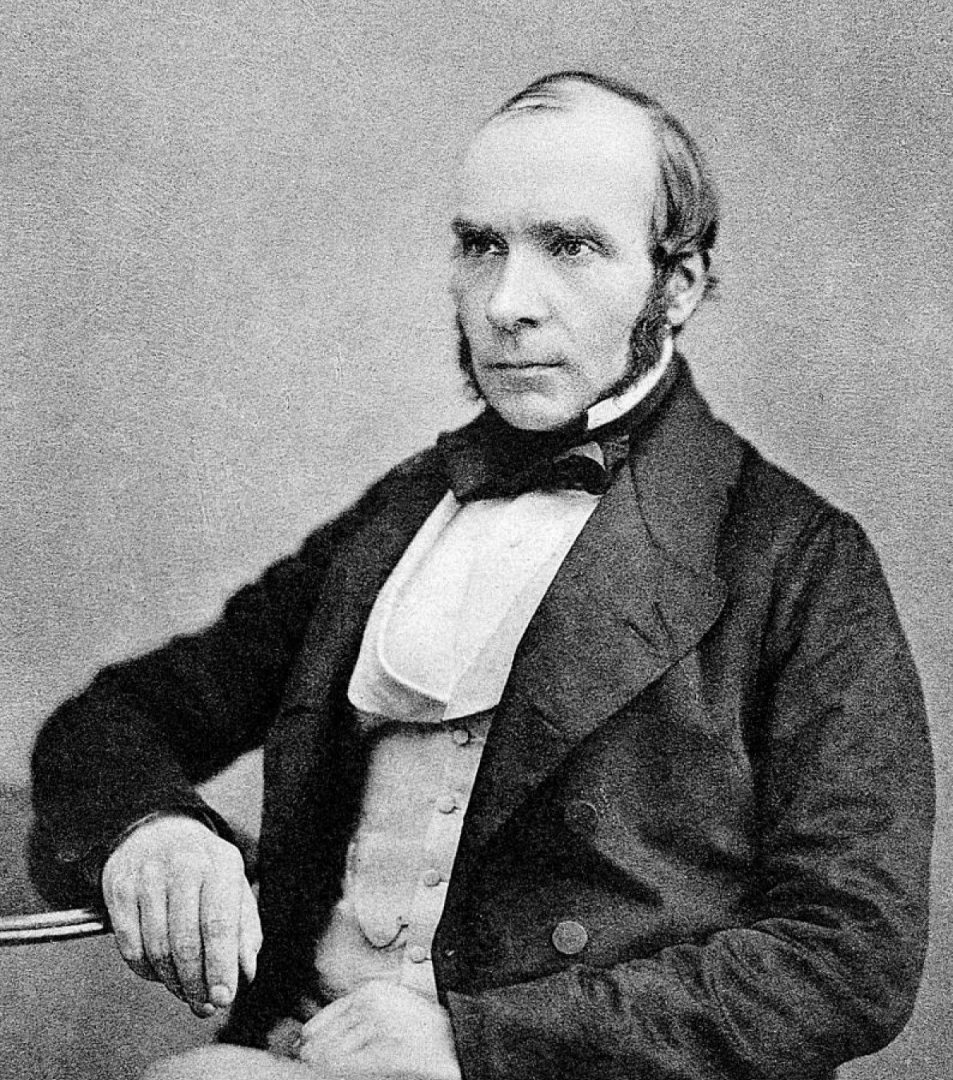


Data Mining y Visualización

Vanessa P. Araya
www.varaya.cl
Equipo ILDA, Inria

El misterio del cólera





John Snow: **1813–1858**

Médico inglés, considerado padre de la epidemiología moderna

Recolección de datos

TABLE VI.

The Mortality from Cholera in 1854, in Thirty-one Sub-Districts, as compared with Calculations founded on the Results shown in Table v.

| ion Districts. | Registration Sub-Districts. | Population in 1851. | Estimated population supplied with water as under. | | | Deaths from cholera in 1854. | | Calculated mortality in the population supplied with water as under. | | | |
|----------------|-----------------------------|---------------------|--|-------------|--------------------------|------------------------------|---------------------------|--|-------------------------------|--------------------|-------------|
| | | | Southwark and Vauxhall Co. | Lambeth Co. | Both Companies together. | Total deaths. | Deaths per 10,000 living. | Southwark and Vauxhall Co. at 160 per 10,000. | Lambeth Co. at 27 per 10,000. | The two Companies. | Calculated. |
| r, Southw. | 1. Christchurch | 10,022 | 2,015 | 13,234 | 16,149 | 113 | 71 | 46 | 36 | 82 | |
| | 2. St. Saviour | 10,700 | 16,337 | 898 | 17,235 | 378 | 192 | 261 | 2 | 263 | |
| - - - - | 1. St. Olave | 8,015 | 8,745 | 0 | 8,745 | 161 | 201 | 140 | 0 | 140 | |
| | 2. St. John, Horselydown | 11,300 | 9,300 | 0 | 9,300 | 152 | 134 | 150 | 0 | 150 | |
| ey - - - | 1. St. James | 18,899 | 23,173 | 603 | 23,866 | 362 | 192 | 370 | 2 | 372 | |
| | 2. St. Mary Magdalen . | 13,934 | 17,258 | 0 | 17,258 | 247 | 177 | 276 | 0 | 276 | |
| s, Southw. | 3. Leather Market . . . | 15,295 | 14,003 | 1,092 | 15,095 | 237 | 155 | 224 | 3 | 227 | |
| | 1. Kent Road | 18,126 | 12,630 | 3,997 | 16,627 | 177 | 98 | 202 | 11 | 213 | |
| - - - - | 2. Borough Road . . . | 15,862 | 8,937 | 6,672 | 15,609 | 271 | 171 | 143 | 18 | 161 | |
| | 3. London Road . . . | 17,836 | 2,872 | 11,407 | 14,369 | 95 | 53 | 46 | 31 | 79 | |
| n - - - - | 1. Trinity | 20,922 | 10,132 | 8,370 | 18,502 | 211 | 101 | 162 | 22 | 184 | |
| | 2. St. Peter, Walworth . | 29,861 | 14,274 | 10,724 | 24,998 | 391 | 131 | 228 | 29 | 257 | |
| | 3. St. Mary | 14,033 | 2,983 | 5,484 | 8,467 | 92 | 66 | 48 | 15 | 63 | |

Visualización de los datos



Conclusiones



Descubrimiento del cólera

**NOT TO DRINK ANY WATER
WHICH HAS NOT
PREVIOUSLY BEEN BOILED.**

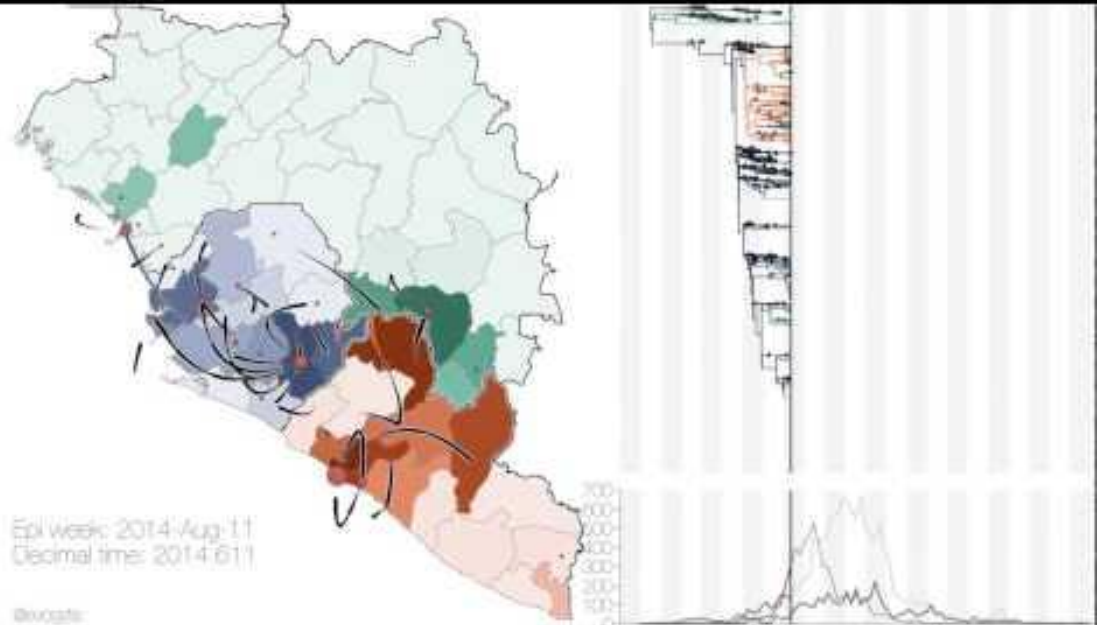
Fresh Water ought to be Boiled every Morning for the day's use, and what remains of it ought to be thrown away at night. The Water ought not to stand where any kind of dirt can get into it, and great care ought to be given to see that Water Butts and Cisterns are free from dirt.



Gytis Dudas et al.

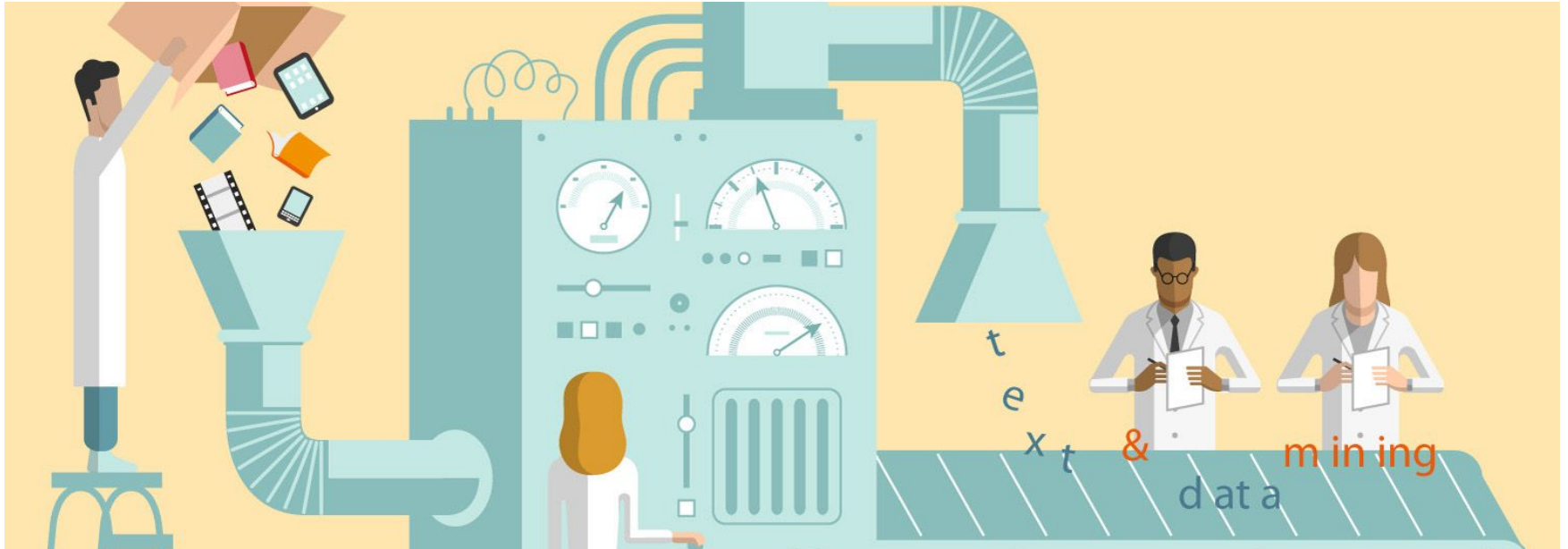
**Virus genomes reveal
factors that spread
and sustained the
Ebola epidemic.**

Nature, 2017



¿Qué es minería de datos?

Descubrir automáticamente información útil en grandes repositorios de datos



Tareas a seguir

- Establecer el problema
- Obtener datos adecuados
- Limpiar y explorar los datos -> ¡Visualización!
- Modelar nuestros datos
- Interpretar y comunicar

Tareas a seguir

- Establecer el problema
- Obtener datos adecuados
- Limpiar y explorar los datos -> ¡Visualización!
- Modelar nuestros datos
- Interpretar y comunicar

Parte 1:

Preprocesamiento y Visualización

- Establecer el problema

- Obtener datos adecuados

Parte 2:

Aprendizaje Supervisado

- Limpiar y explorar los datos -> ¡Visualización!

- Modelar nuestros datos

Parte 3:

Aprendizaje No Supervisado

Estructura general

Parte 1:

Preprocesamiento y Visualización

Parte 2:

Aprendizaje Supervisado

Parte 3:

Aprendizaje No Supervisado

Estructura general

Parte 1:

Preprocesamiento y Visualización

Parte 2:

Aprendizaje Supervisado

Parte 3:

Aprendizaje No Supervisado

Establecer el problema

- ¿Qué enfermedad quiero entender?
- ¿Por qué entender esta enfermedad es importante?
- ¿Quiénes se ven afectados por esta enfermedad?
- ¿Qué espero obtener de mi análisis que pueda cambiar?

Obtener los datos puede ser difícil

As far as public datasets, the public has to realize that you're not going to be able to see the raw information. Just imagine the sorts of privacy rights issues you'd have in the U.S. with HIPAA compliance. Public health officials are not going to share exact names and identities of those affected, never mind all their suspected contacts over weeks. The best that the public can hope for will be analysis of aggregate, anecdotal or incidental data and metadata—such as social media posts, or global flight information. Some companies like BlueDot and Metabiota already have services that provide analysis of such data. But, again, these are not public open data sets.

Fuente <https://towardsdatascience.com/chasing-the-data-coronavirus-802d8a1c4e9a>

Obtener los datos puede ser difícil



Fuente <https://www.minciencia.gob.cl/covid19>

Algunas fuentes de datos

- <https://archive.ics.uci.edu/ml/index.php>
- <https://www.kaggle.com/>
- <https://www.who.int/>
- <https://github.com/curran/data>

Preprocesar los datos

- Creación de atributos
- Selección de un subconjunto de atributos
- Agregación
- Normalización
- Muestreo
- Reducción de dimensionalidad
- Discretización y binarización
- Transformación

Lectura recomendada:

<https://jakevdp.github.io/PythonDataScienceHandbook/05.04-feature-engineering.html>

Explorar los datos usando Visualización

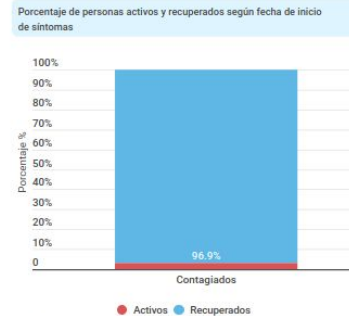
¿Por qué?

- La visión humana es muy poderosa
- Tenemos la capacidad de reconocer patrones

-> Colaboración entre el humano y el computador

Explorar los datos usando Visualización

Porcentaje de activos y recuperados según fecha de inicio de síntomas



Download data

Fuente: Base de datos Ministerio de Ciencia, en base a Reporte Diario Coronavirus Ministerio de Salud.

Casos confirmados COVID-19, según sintomatología reportada



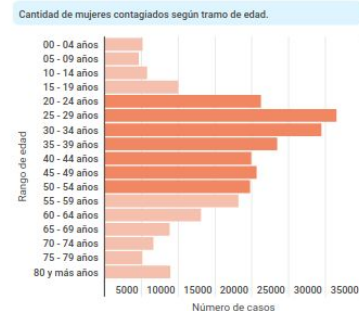
Download data

Fuente: Información proviene del informe de situación covid-19 (<http://epi.minsal.cl/informes-covid-19/>)

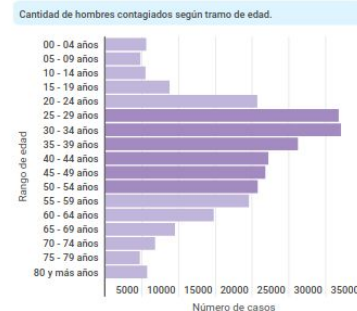
Número total de casos confirmados según tramo de edad



Mujeres



Hombres



Fuente <https://www.gob.cl/coronavirus/cifrasoficiales/#datos>

Cuarteto de Anscombe

Anscombe's quartet

| I | | II | | III | | IV | |
|------|-------|------|------|------|-------|------|-------|
| x | y | x | y | x | y | x | y |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

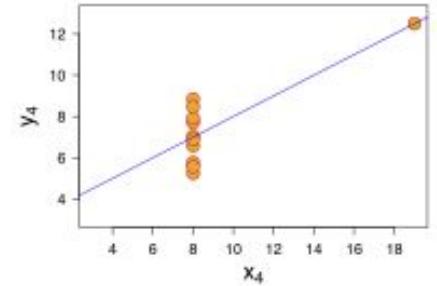
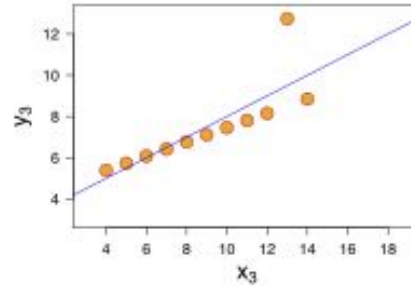
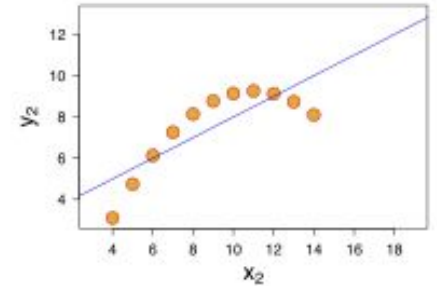
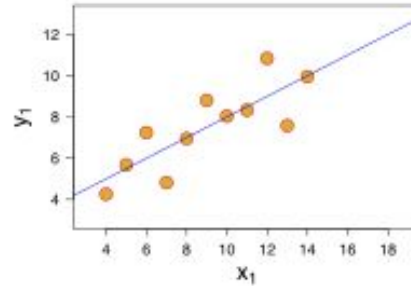
| Propiedad | Valor | Exactitud |
|-----------------------------|-------|---------------------|
| Promedio de X | 9 | Exacto |
| Varianza de la muestra de x | 11 | Exacto |
| Promedio de y | 5.50 | Hasta dos decimales |
| Varianza de la muestra de y | 4.125 | ± 0.003 |
| Correlación entre x ey | 0.816 | Hasta 3 decimales |

Fuente: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Cuarteto de Anscombe

Anscombe's quartet

| I | | II | | III | | IV | |
|------|-------|------|------|------|-------|------|-------|
| x | y | x | y | x | y | x | y |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

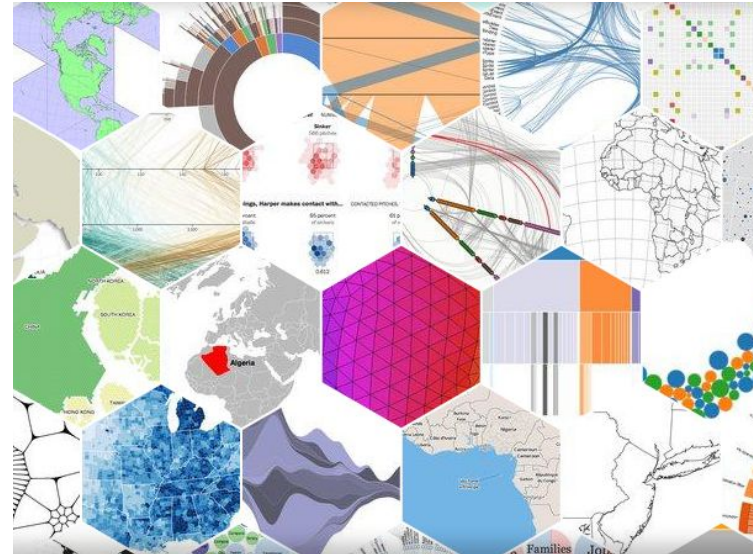


Algunas herramientas de visualización disponibles

Tableau, pagada :'

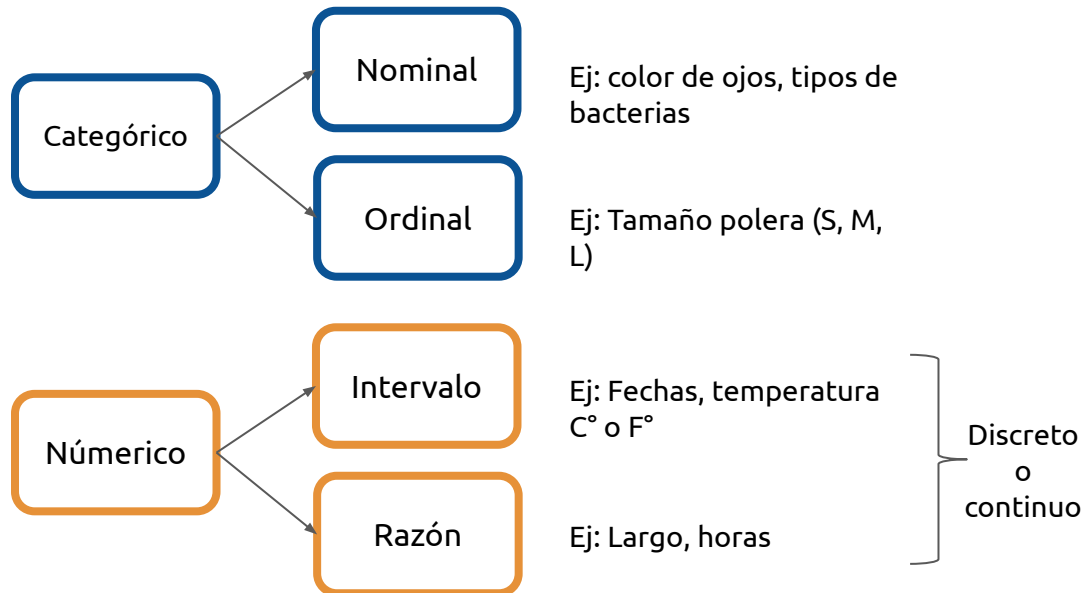


D3js (web interactivas)



¿Cómo visualizamos de manera efectiva?

1) Entender los datos



¿Cómo visualizamos de manera efectiva?

2) Ver qué posibilidades tengo

➔ Position

➔ Horizontal



➔ Vertical



➔ Both



➔ Color



➔ Shape



➔ Tilt



➔ Size

➔ Length



➔ Area



➔ Volume



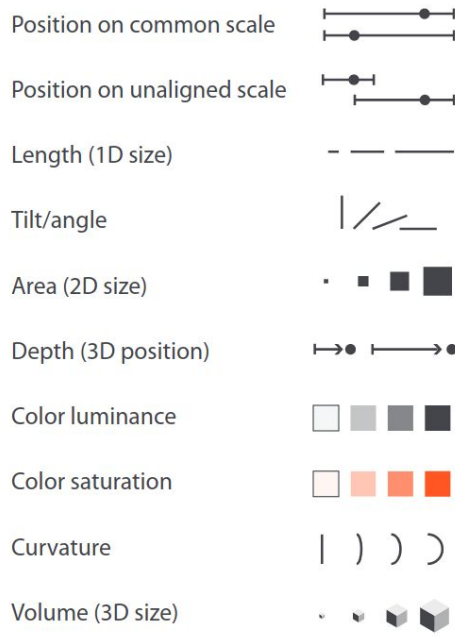
Visual (encoding) channels control the appearance of marks

¿Cómo visualizamos de manera efectiva?

3) Buscar mejor match

Channels: Expressiveness Types and Effectiveness Ranks

➔ **Magnitude Channels: Ordered Attributes**

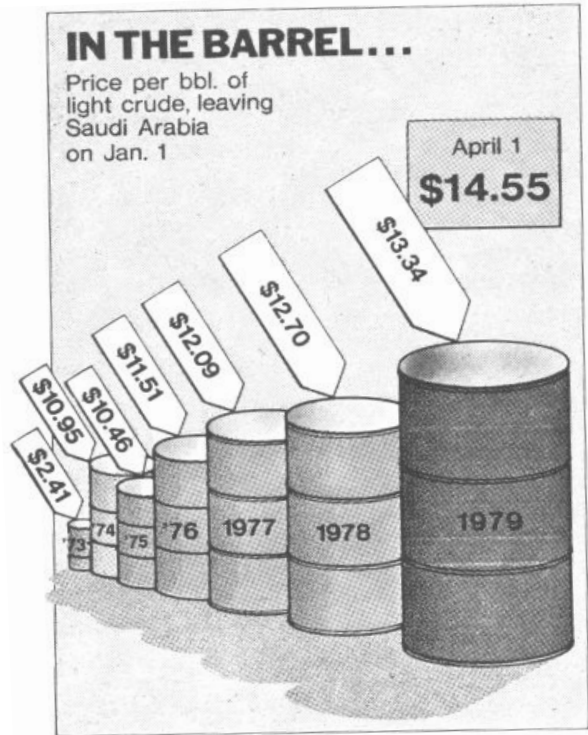


➔ **Identity Channels: Categorical Attributes**



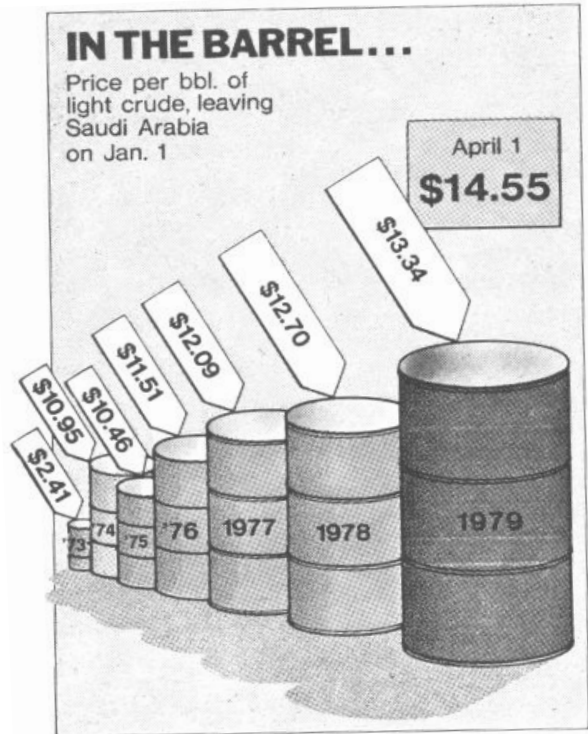
Fuente: Visualization Analysis and Design, Tamara Munzner

Por ejemplo

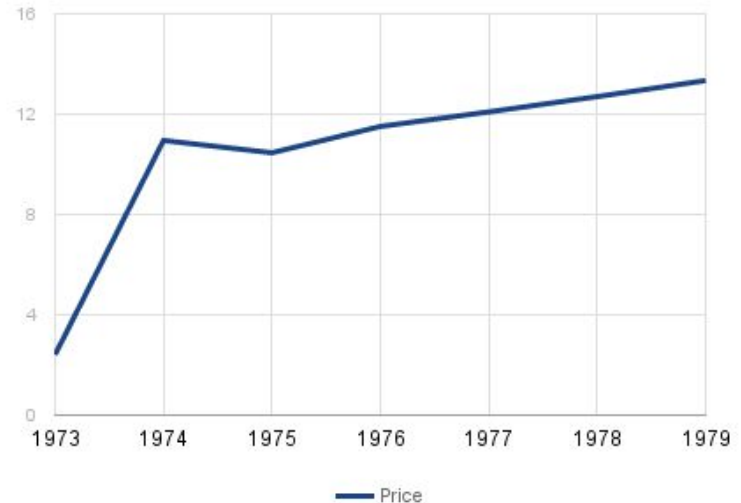


Fuente: The Visual Display of Quantitative Information - Edward R. Tufte

Por ejemplo

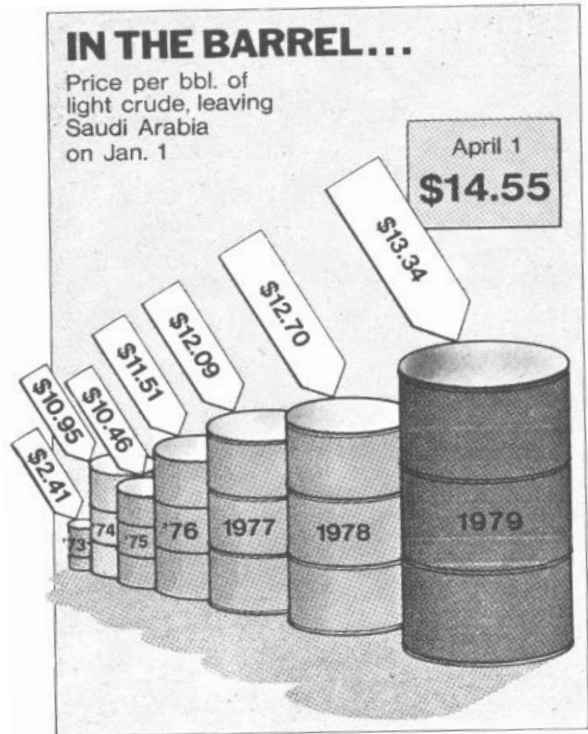


Price per bbl. of light crude leaving Saudi Arabia



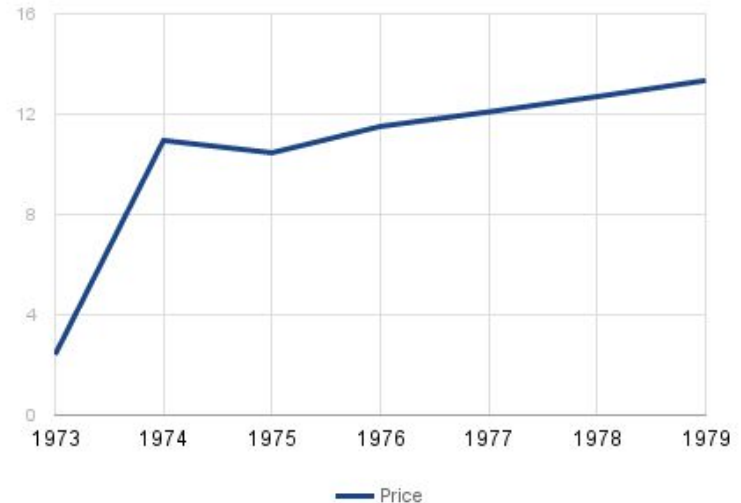
Fuente: The Visual Display of Quantitative Information - Edward R. Tufte

Por ejemplo



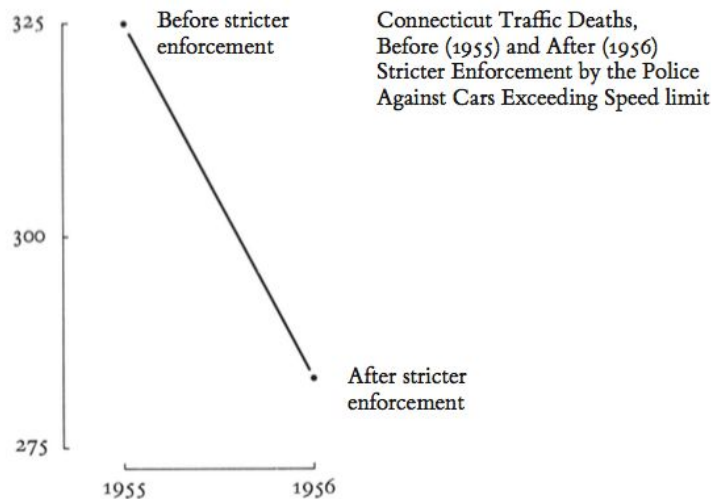
Número de dimensiones en una vis no debe exceder el n° de dimensiones de los datos:

Price per bbl. of light crude leaving Saudi Arabia

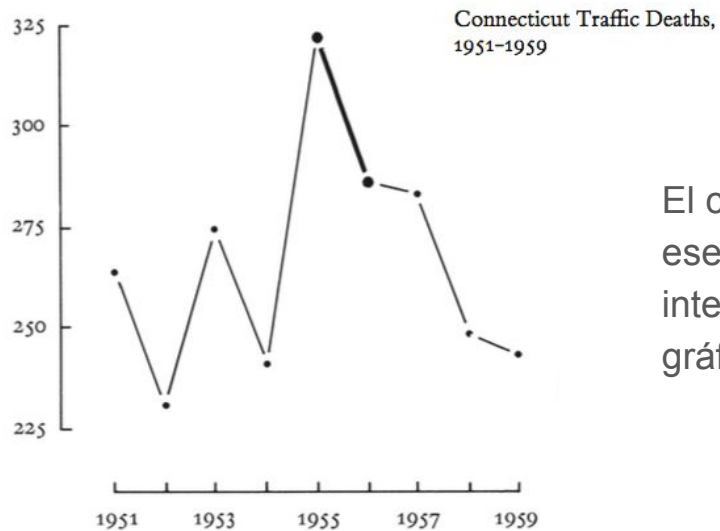
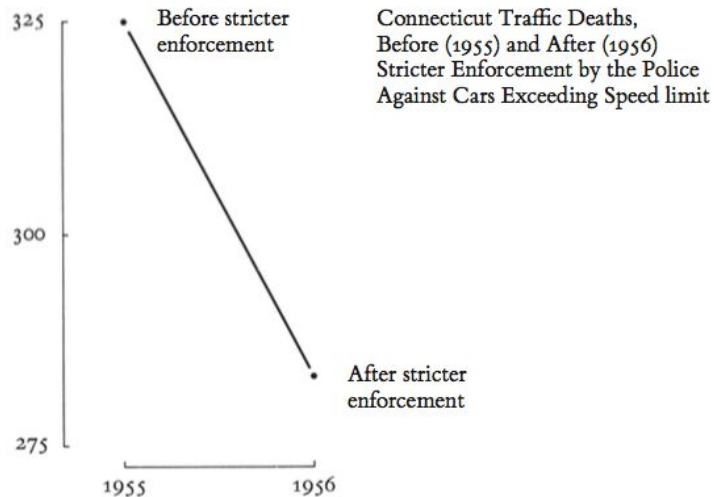


Fuente: The Visual Display of Quantitative Information - Edward R. Tufte

Evitar sesgos



Evitar sesgos



El contexto es esencial para la integridad de un gráfico

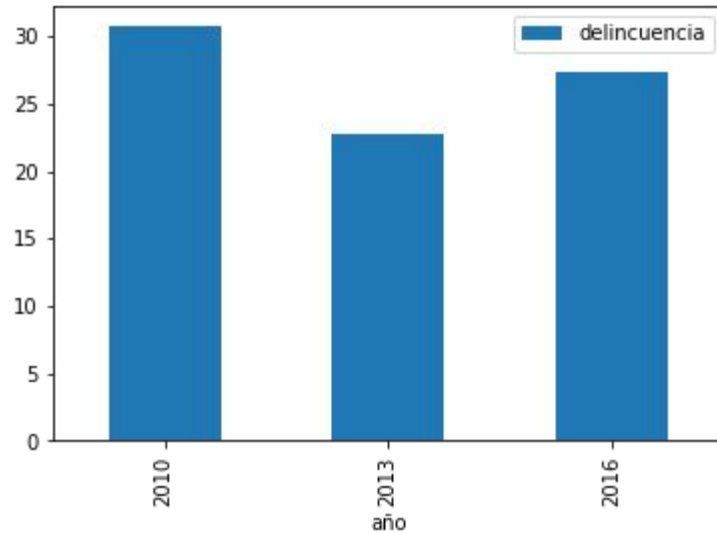
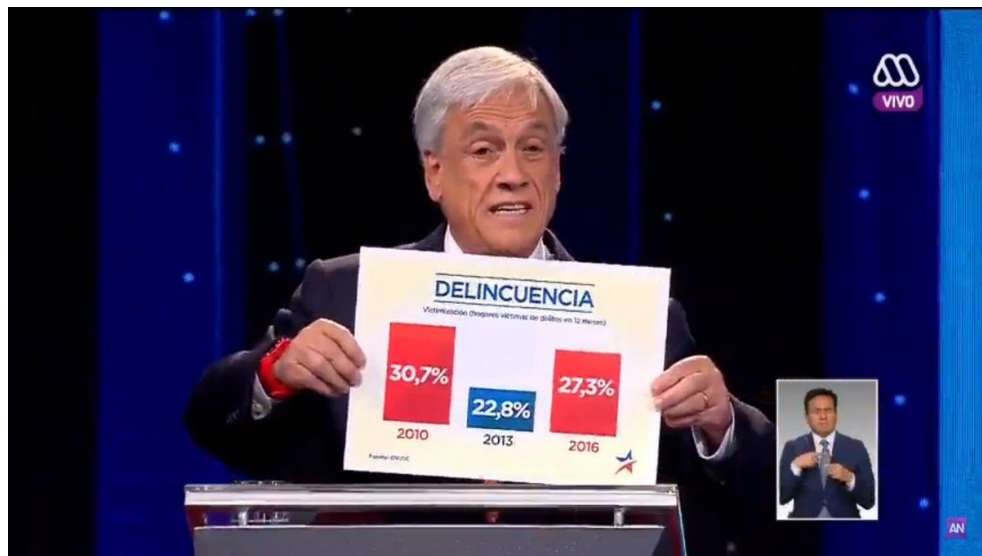
Evitar sesgos



Fuente:

<https://viz.wtf/post/167382195183/the-former-president-of-chile-and-now-candidate>

Evitar sesgos



Fuente:

<https://viz.wtf/post/167382195183/the-former-president-of-chile-and-now-candidate>

Recursos acerca de visualización

- Slides de Tamara Munzner: <https://www.cs.ubc.ca/~tmm/vadbook/>
- Edward Tufte: <https://www.edwardtufte.com/tufte/>, en particular el libro The Visual Display of Quantitative Information
- Como no hacer visualizaciones: <https://viz.wtf/>

Resumen

- Establecer el problema
- Obtener datos adecuados
- Limpiar y explorar los datos ->
¡Visualización!
- Modelar nuestros datos

Resumen

- Establecer el problema
- Obtener datos adecuados
- Limpiar y explorar los datos -> ¡Visualización!
- Modelar nuestros datos

-> Ir a Jupyter lab parte 1

Estructura general

Parte 1:

Preprocesamiento y Visualización

Parte 2:

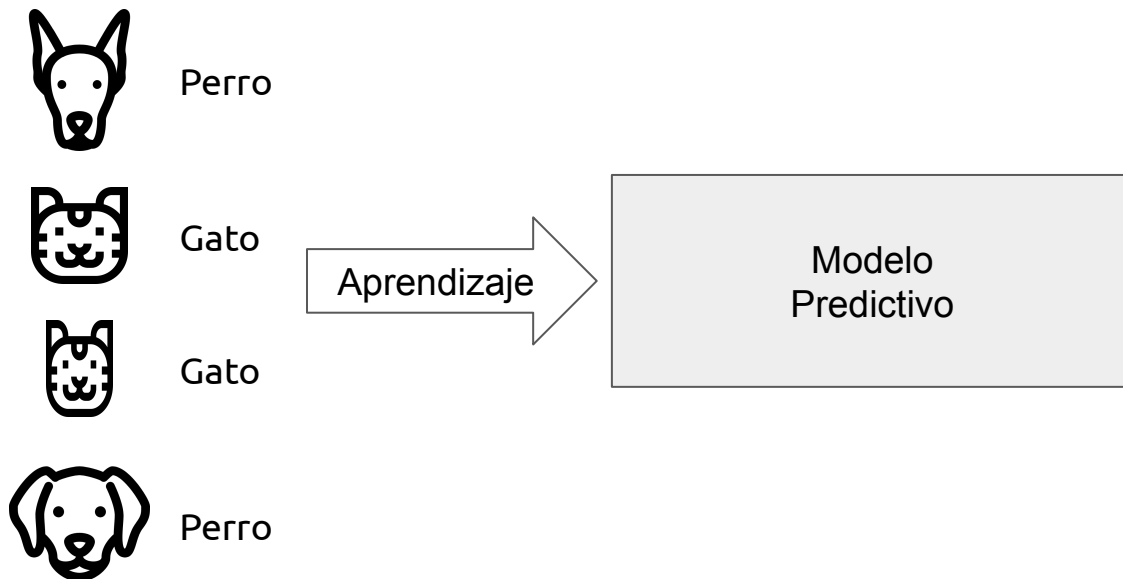
Aprendizaje Supervisado

Parte 3:

Aprendizaje No Supervisado

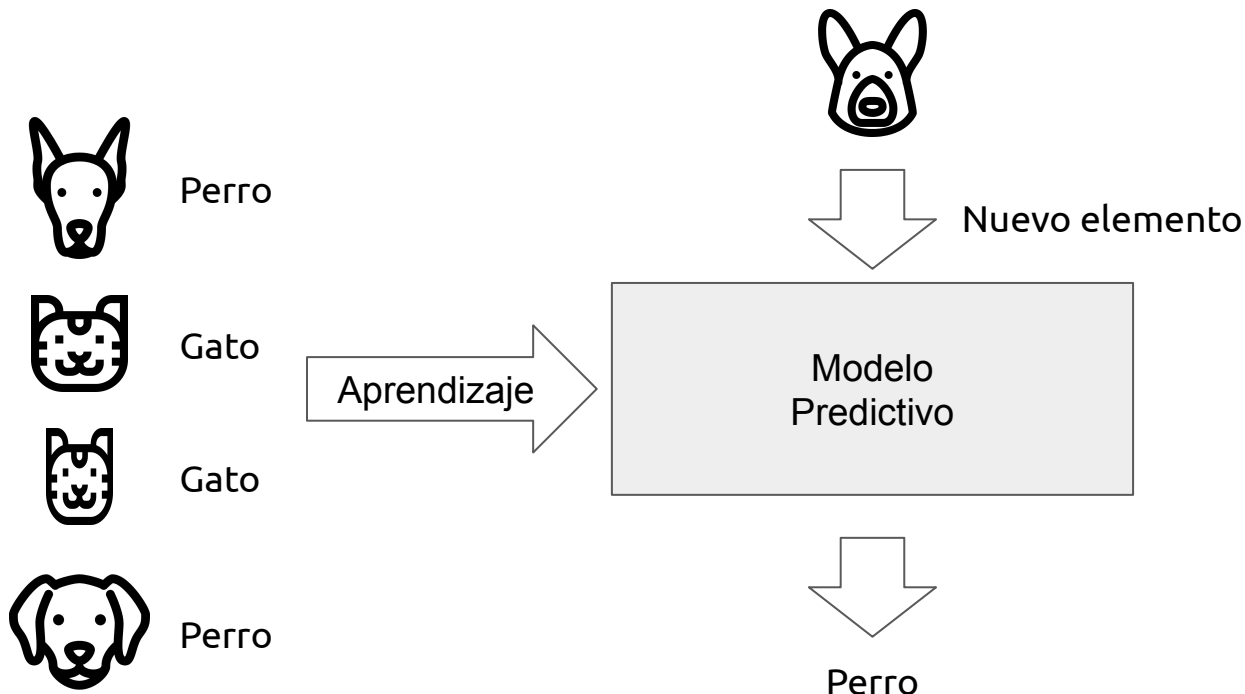
Aprendizaje Supervisado

Crear una función
que dado un set de
datos de
entrenamiento,
pueda predecir el
output deseado



Aprendizaje Supervisado

Crear una función
que dado un set de
datos de
entrenamiento,
pueda predecir el
output deseado



Dos tipos de aprendizaje supervisado

Clasificación:

Queremos obtener valores discretos.

Por ejemplo:

- Predecir si una célula cancerígena es maligna o benigna
- Predecir si un email es spam

Dos tipos de aprendizaje supervisado

Clasificación:

Queremos obtener valores discretos.

Por ejemplo:

- Predecir si una célula cancerígena es maligna o benigna
- Predecir si un email es spam

Regresión:

Nuestro target es continuo.

Por ejemplo: predecir el nivel de azúcar en la sangre dado una dosis de una medicina.

Podemos crear una función:

$$\text{Nivel_azucar(dosis)} = \text{nivel_azucar_base} + \text{factor} * \text{dosis}$$

Ingredientes necesarios

- Datos de entrenamiento
- Datos a predecir
- Método a usar

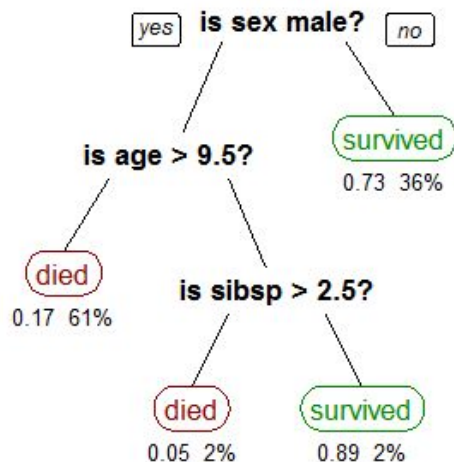
¿Qué obtenemos?

- Modelo
- Métricas de evaluación

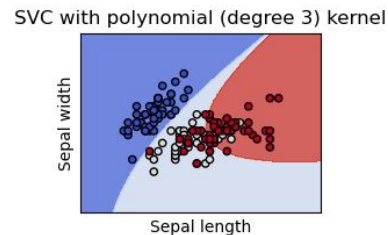
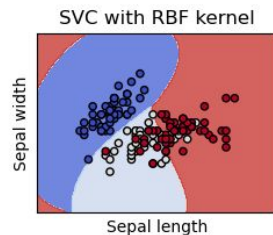
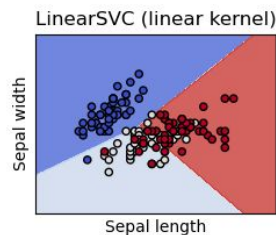
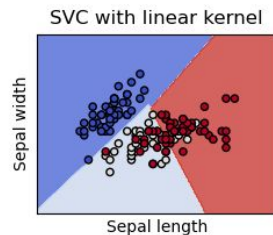


Nosotros veremos dos métodos de clasificación

Árboles de decisión

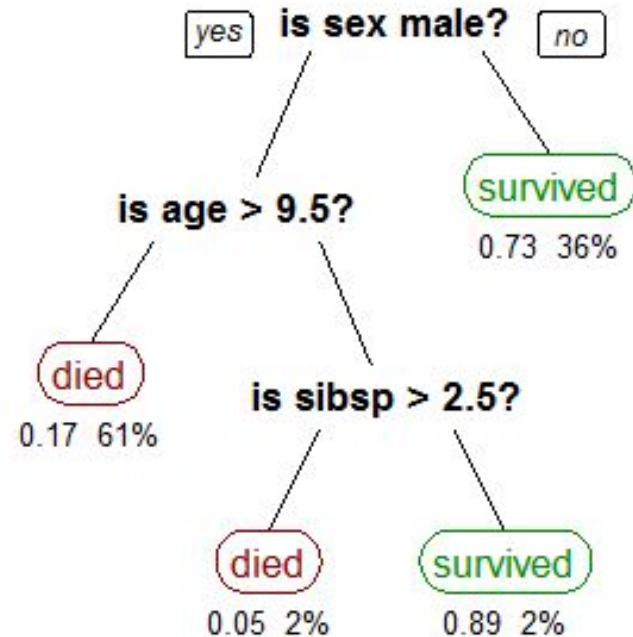


Support Vector Machine



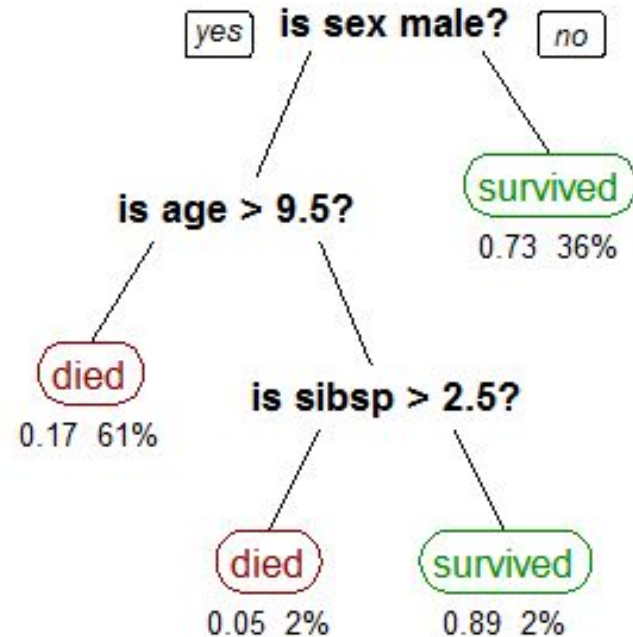
Árboles de decisión

Un método no paramétrico que construye un modelo aprendiendo reglas de decisión.



Árboles de decisión

Un método **no paramétrico** que construye un modelo aprendiendo reglas de decisión.



Métodos paramétricos

Asumen algunos parámetros de entrada.

¿Por qué? Porque hacen la vida más simple.

Funcionan así:

- Seleccionan la forma de una función
- Aprenden los coeficientes de esa función dado los datos de entrada

Métodos paramétricos

Asumen algunos parámetros de entrada.

¿Por qué? Porque hacen la vida más simple.

Funcionan así:

- Seleccionan la forma de una función
- Aprenden los coeficientes de esa función dado los datos de entrada

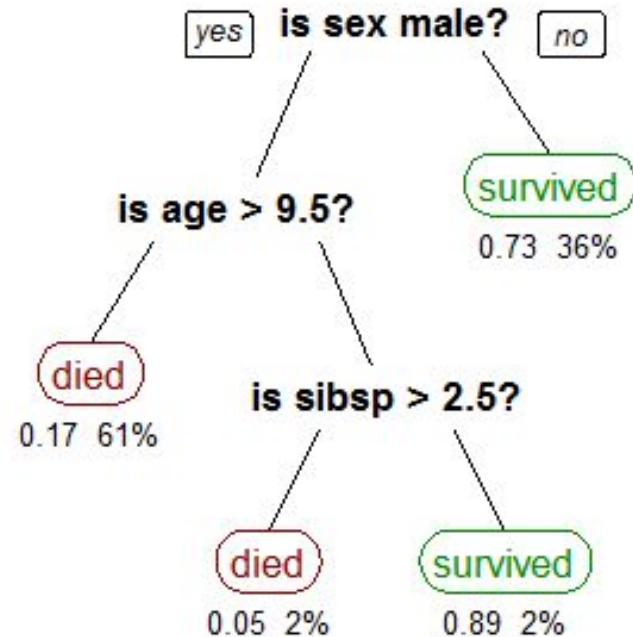
Métodos no paramétricos

No asumen nada y pueden aprender más libremente.

¿Por qué? Porque a veces no sabemos nada de los datos y no queremos tener que definir ningún parámetro a priori

Árboles de decisión

Un método **no paramétrico** que construye un modelo aprendiendo reglas de decisión.



Algoritmo CART

| Sexo | Fuma | Diabetes | Muerte |
|------|------|----------|--------|
| F | Sí | Sí | Sí |
| F | No | Sí | No |
| M | No | No | Sí |
| M | No | Sí | No |
| M | Sí | No | Sí |
| F | Sí | No | Sí |
| M | No | No | No |
| M | Sí | Sí | Sí |
| F | Sí | Sí | Sí |
| F | Sí | Sí | Sí |

Algoritmo CART

| Sexo | Fuma | Diabetes | Muerte |
|------|------|----------|--------|
| F | Sí | Sí | Sí |
| F | No | Sí | No |
| M | No | No | Sí |
| M | No | Sí | No |
| M | Sí | No | Sí |
| F | Sí | No | Sí |
| M | No | No | No |
| M | Sí | Sí | Sí |
| F | Sí | Sí | Sí |
| F | Sí | Sí | Sí |

¿Qué atributo me da más información para cortar el árbol en dos?

El índice Gini mide el grado de que una variable haya sido clasificada de manera errónea.

$$\text{Gini} = 1 - \sum_{i=1}^n (p_i)^2$$

Algoritmo CART

| Sexo | Fuma | Diabetes | Muerte |
|------|------|----------|--------|
| F | Sí | Sí | Sí |
| F | No | Sí | No |
| M | No | No | Sí |
| M | No | Sí | No |
| M | Sí | No | Sí |
| F | Sí | No | Sí |
| M | No | No | No |
| M | Sí | Sí | Sí |
| F | Sí | Sí | Sí |
| F | Sí | Sí | Sí |

¿Qué atributo me da más información para cortar el árbol en dos?

El índice Gini mide el grado de que una variable haya sido clasificada de manera errónea.

Si es igual a cero, entonces todos mis elementos están bien clasificados. Si es igual a 1, entonces mis elementos están asignados aleatoriamente

$$\text{Gini} = 1 - \sum_{i=1}^n (p_i)^2$$

Algoritmo CART

| Sexo | Fuma | Diabetes | Muerte |
|------|------|----------|--------|
| F | Sí | Sí | Sí |
| F | No | Sí | No |
| M | No | No | Sí |
| M | No | Sí | No |
| M | Sí | No | Sí |
| F | Sí | No | Sí |
| M | No | No | No |
| M | Sí | Sí | Sí |
| F | Sí | Sí | Sí |
| F | Sí | Sí | Sí |

$$\text{Gini} = 1 - \sum_{i=1}^n (p_i)^2$$

Calculemos el Gini index para el atributo sexo:

$$P(\text{sexo} = F \text{ \& muerte} = \text{Sí}) = 4/5$$

$$P(\text{sexo} = F \text{ \& muerte} = \text{No}) = 1/5$$

$$Gi_{\text{sexo}_F} = 1 - ((4/5)^2 + (1/5)^2) = 0.32$$

$$P(\text{sexo} = M \text{ \& muerte} = \text{Sí}) = 3/5$$

$$P(\text{sexo} = M \text{ \& muerte} = \text{No}) = 2/5$$

$$Gi_{\text{sexo}_M} = 1 - ((3/5)^2 + (2/5)^2) = 0.48$$

$$Gi_{\text{sexo}} = P(\text{sexo}==F) * 0.32 + P(\text{sexo}==M) * 0.48 = 0.4$$

$$Gi_{\text{sexo}} = 5/10 * 0.32 + 5/10 * 0.48 = 0.4$$

Algoritmo CART

| Sexo | Fuma | Diabetes | Muerte |
|------|------|----------|--------|
| F | Sí | Sí | Sí |
| F | No | Sí | No |
| M | No | No | Sí |
| M | No | Sí | No |
| M | Sí | No | Sí |
| F | Sí | No | Sí |
| M | No | No | No |
| M | Sí | Sí | Sí |
| F | Sí | Sí | Sí |
| F | Sí | Sí | Sí |

$$\text{Gini} = 1 - \sum_{i=1}^n (p_i)^2$$

Calculemos el Gini index para el atributo Fuma:

$$P(\text{fuma} = \text{Sí} \ \& \ \text{muerte} = \text{Sí}) = 6/6$$

$$P(\text{fuma} = \text{Sí} \ \& \ \text{muerte} = \text{No}) = 0/6$$

$$G_{i_fuma_Si} = 1 - ((1)^2 + (0)^2) = 0$$

$$P(\text{fuma} = \text{No} \ \& \ \text{muerte} = \text{Sí}) = 1/4$$

$$P(\text{fuma} = \text{No} \ \& \ \text{muerte} = \text{No}) = 3/4$$

$$G_{i_fuma_no} = 1 - ((1/4)^2 + (3/4)^2) = 0.375$$

$$G_{i_fuma} = 6/10 * 0 + 4/10 * 0.375 = 0.15$$

Algoritmo CART

| Sexo | Fuma | Diabetes | Muerte |
|------|------|----------|--------|
| F | Sí | Sí | Sí |
| F | No | Sí | No |
| M | No | No | Sí |
| M | No | Sí | No |
| M | Sí | No | Sí |
| F | Sí | No | Sí |
| M | No | No | No |
| M | Sí | Sí | Sí |
| F | Sí | Sí | Sí |
| F | Sí | Sí | Sí |

$$\text{Gini} = 1 - \sum_{i=1}^n (p_i)^2$$

Calculemos el Gini index para el atributo Diabetes:

$$\begin{aligned} P(\text{diabetes} = \text{Sí} \ \& \ \text{muerte} = \text{Sí}) &= 4/6 \\ P(\text{diabetes} = \text{Sí} \ \& \ \text{muerte} = \text{No}) &= 2/6 \end{aligned}$$

$$\text{Gi_diabetes_Si} = 1 - ((4/6)^2 + (2/6)^2) = 0.44$$

$$\begin{aligned} P(\text{diabetes} = \text{No} \ \& \ \text{muerte} = \text{Sí}) &= 3/4 \\ P(\text{diabetes} = \text{No} \ \& \ \text{muerte} = \text{No}) &= 1/4 \end{aligned}$$

$$\text{Gi_diabetes_no} = 1 - ((3/4)^2 + (1/4)^2) = 0.375$$

$$\text{Gi_diabetes} = 6/10 * 0.44 + 4/10 * 0.375 = 0.414$$

Algoritmo CART

| Sexo | Fuma | Diabetes | Muerte |
|------|------|----------|--------|
| F | Sí | Sí | Sí |
| F | No | Sí | No |
| M | No | No | Sí |
| M | No | Sí | No |
| M | Sí | No | Sí |
| F | Sí | No | Sí |
| M | No | No | No |
| M | Sí | Sí | Sí |
| F | Sí | Sí | Sí |
| F | Sí | Sí | Sí |

$$\text{Gini} = 1 - \sum_{i=1}^n (p_i)^2$$

$G_i_{\text{sexo}} = 0.4$

$G_i_{\text{fuma}} = 0.15$

$G_i_{\text{diabetes}} = 0.414$

Algoritmo CART

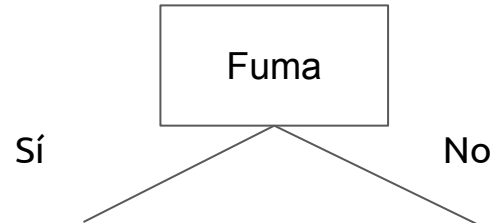
| Sexo | Fuma | Diabetes | Muerte |
|------|------|----------|--------|
| F | Sí | Sí | Sí |
| F | No | Sí | No |
| M | No | No | Sí |
| M | No | Sí | No |
| M | Sí | No | Sí |
| F | Sí | No | Sí |
| M | No | No | No |
| M | Sí | Sí | Sí |
| F | Sí | Sí | Sí |
| F | Sí | Sí | Sí |

$$\text{Gini} = 1 - \sum_{i=1}^n (p_i)^2$$

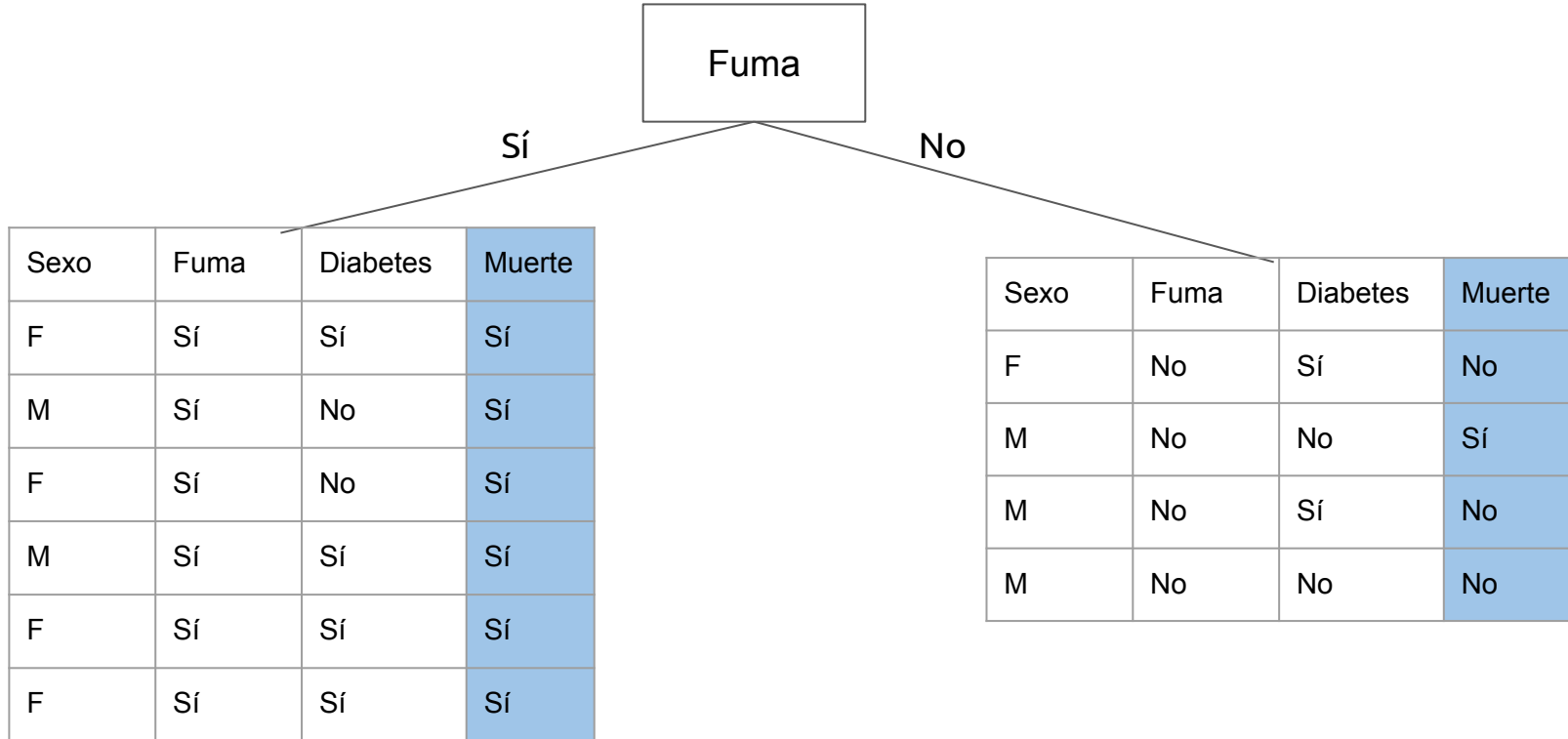
$G_i_{\text{sexo}} = 0.4$

$G_i_{\text{fuma}} = 0.15$

$G_i_{\text{diabetes}} = 0.414$



Algoritmo CART



Recursos acerca de Árboles de decisión

Árboles de regresión:

<https://www.datadriveninvestor.com/2020/04/13/how-do-regression-trees-work/>

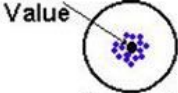
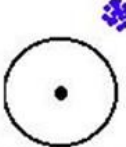
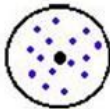
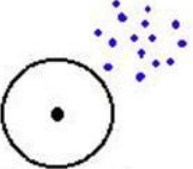
Otros algoritmos:

- ID3: http://saedsayad.com/decision_tree_reg.htm
- C4.5: <https://sefiks.com/2018/05/13/a-step-by-step-c4-5-decision-tree-example/>
- Random forests (usar muchos árboles de decisión):
<https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/>

Clasificación: Cómo saber si nuestro modelo es bueno

- Métricas de desempeño:
 - Accuracy (Exactitud): métrica más usada
 - Error rate (Tasa de error)
- Probando con datos de prueba y otros datasets (generalizable, i.e. bajo error de generalización)

Accuracy (exactitud) y Precision (error)

| www.shmula.com | | Accuracy | |
|----------------|-------------|---|---|
| | | Accurate | Not Accurate |
| Precision | Precise |  <p>Accurate & Precise</p> |  <p>Not Accurate & Precise</p> |
| | Not Precise |  <p>Accurate & Not Precise</p> |  <p>Not Accurate & Not Precise</p> |

Accuracy (exactitud) y Precision (error)

| Clase real | Clase predichas | | |
|------------|-----------------|-------------|------------|
| | | Class = Yes | Class = No |
| | Class = Yes | TP | FN |
| | Class = No | FP | TN |

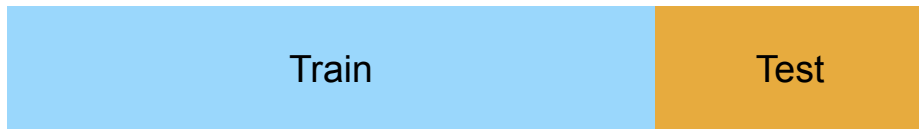
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Probamos con datos de prueba

Dejamos un porcentaje de los datos
para entrenar y otros para testear



Probamos con datos de prueba

Validación cruzada
(cross-validation):
particionar datos en k conjuntos
disjuntos

k-fold:
entrenar con k-1, testear con el
restante

| | | | | | |
|-------------|-------|-------|-------|-------|-------|
| Iteration 1 | Test | Train | Train | Train | Train |
| Iteration 2 | Train | Test | Train | Train | Train |
| Iteration 3 | Train | Train | Test | Train | Train |
| Iteration 4 | Train | Train | Train | Test | Train |
| Iteration 5 | Train | Train | Train | Train | Test |

Fuente: <https://www.mygreatlearning.com/blog/cross-validation/>

Problemas prácticos de clasificación

- Errores de entrenamiento (malos resultados sobre los datos de entrenamiento)
- Errores de generalización (malos resultados sobre datos nuevos)
 - Overfitting
 - Underfitting

Resumen

Modelos de aprendizaje supervisados: toman un conjunto de datos de **entrenamiento** para producir una función que prediga uno o más valores en un dataset nuevo.

Resumen

Modelos de aprendizaje supervisados: toman un conjunto de datos de **entrenamiento** para producir una función que prediga uno o más valores en un dataset nuevo.

Podemos clasificarlos en dos tipos principales: **clasificación** (el objetivo es discreto) o **regresión** (el objetivo es continuo).

Resumen

Modelos de aprendizaje supervisados: toman un conjunto de datos de **entrenamiento** para producir una función que prediga uno o más valores en un dataset nuevo.

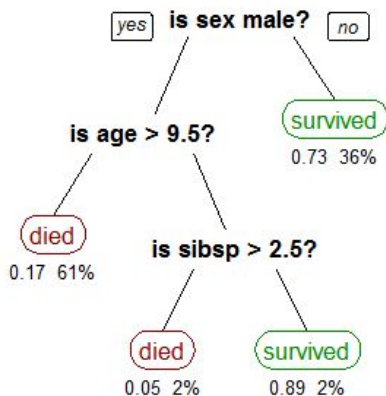
Podemos clasificarlos en dos tipos principales: **clasificación** (el objetivo es discreto) o **regresión** (el objetivo es continuo).

Los métodos pueden ser **paramétricos** (asumen funciones a priori) o **no paramétricos** (no asumen nada).

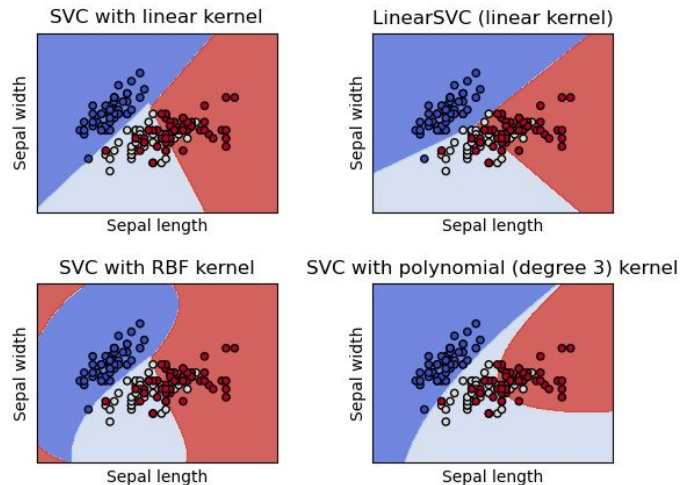
Veremos dos métodos

Árboles de decisión

No paramétrico y que puede ser usado para clasificación o para regresión



Support Vector Machine



-> Ir a Jupyter notebook Parte 2

Estructura general

Parte 1:

Preprocesamiento y Visualización

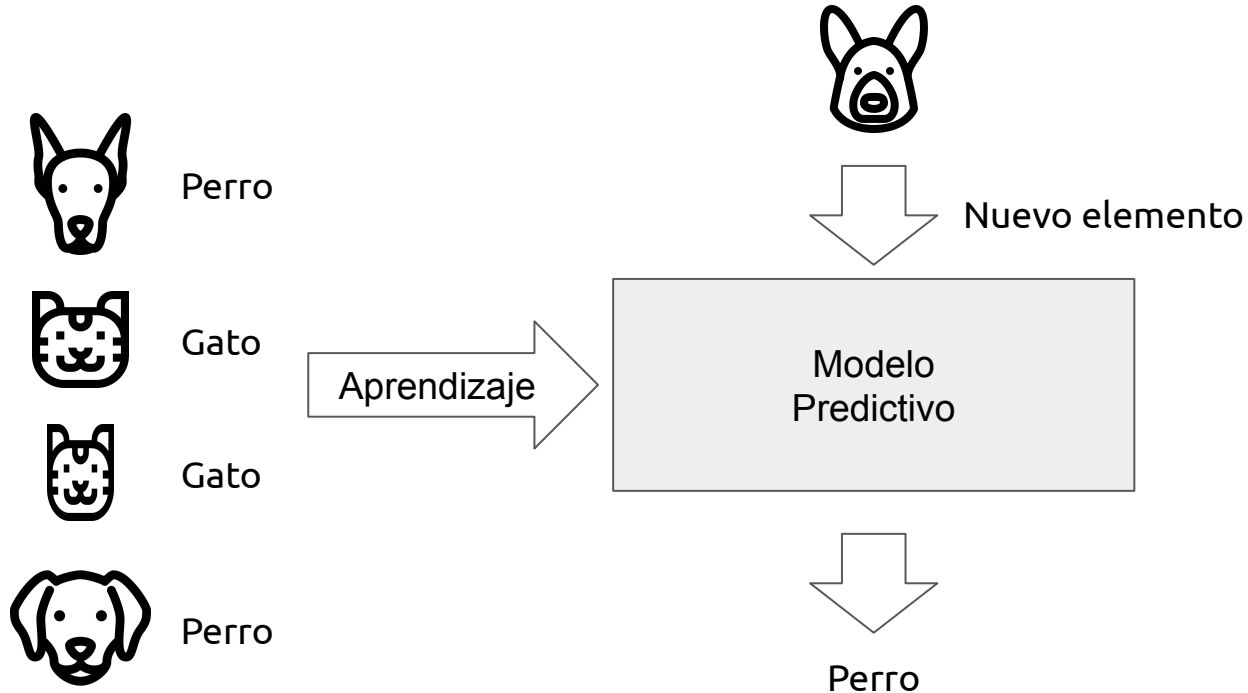
Parte 2:

Aprendizaje Supervisado

Parte 3:

Aprendizaje No Supervisado

Aprendizaje Supervisado



Aprendizaje **No** Supervisado

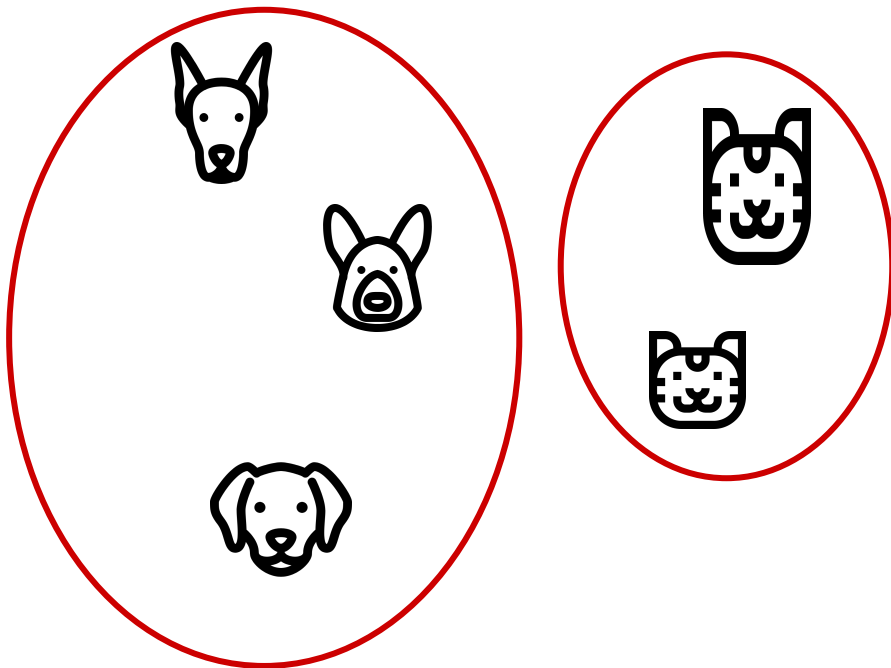
Intenta encontrar
patrones
desconocidos en los
datos, sin
conocimiento previo.



Aprendizaje **No** Supervisado

Intenta encontrar
patrones
desconocidos en los
datos, sin
conocimiento previo.

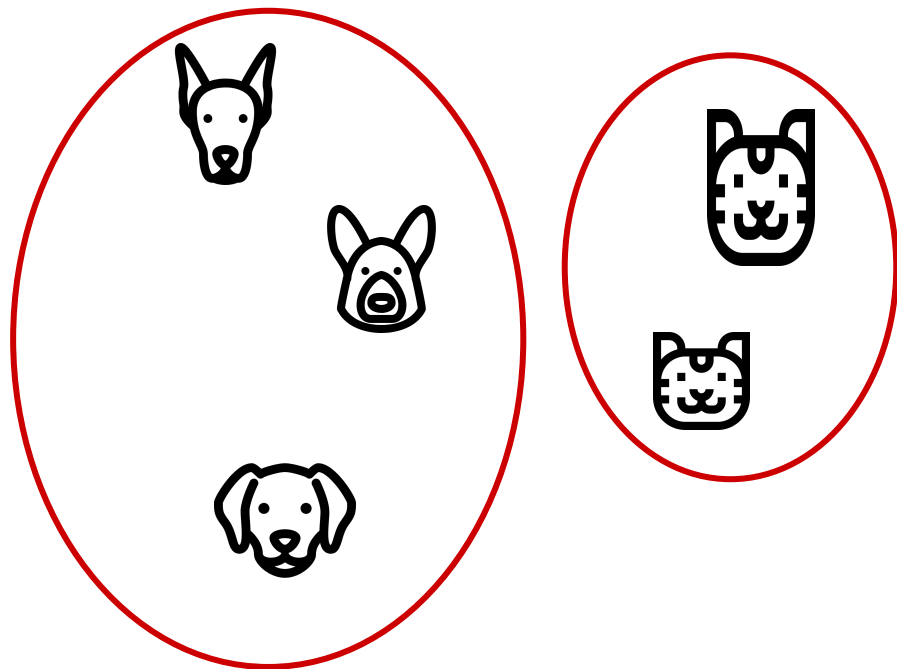
Clustering es uno de
los más conocidos



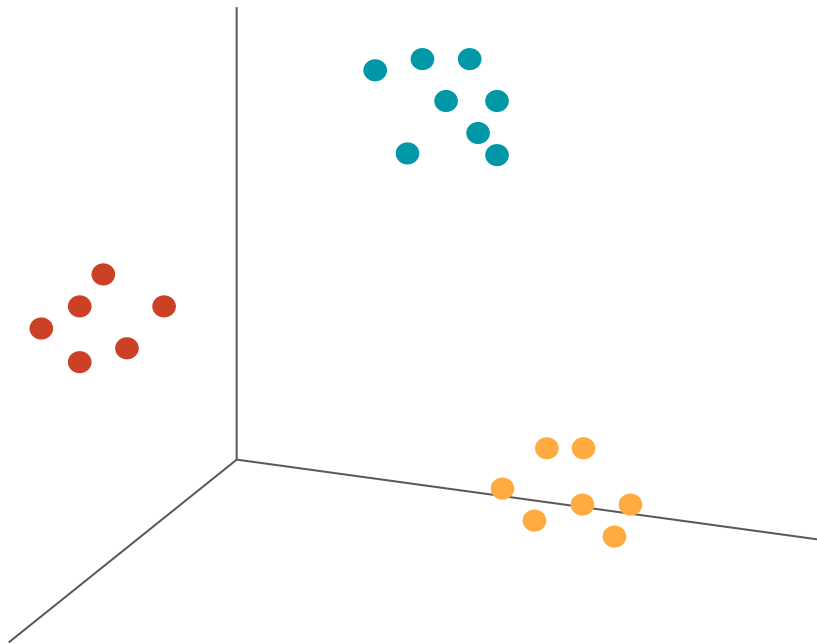
Clustering

Encontrar grupos de objetos especificando que:

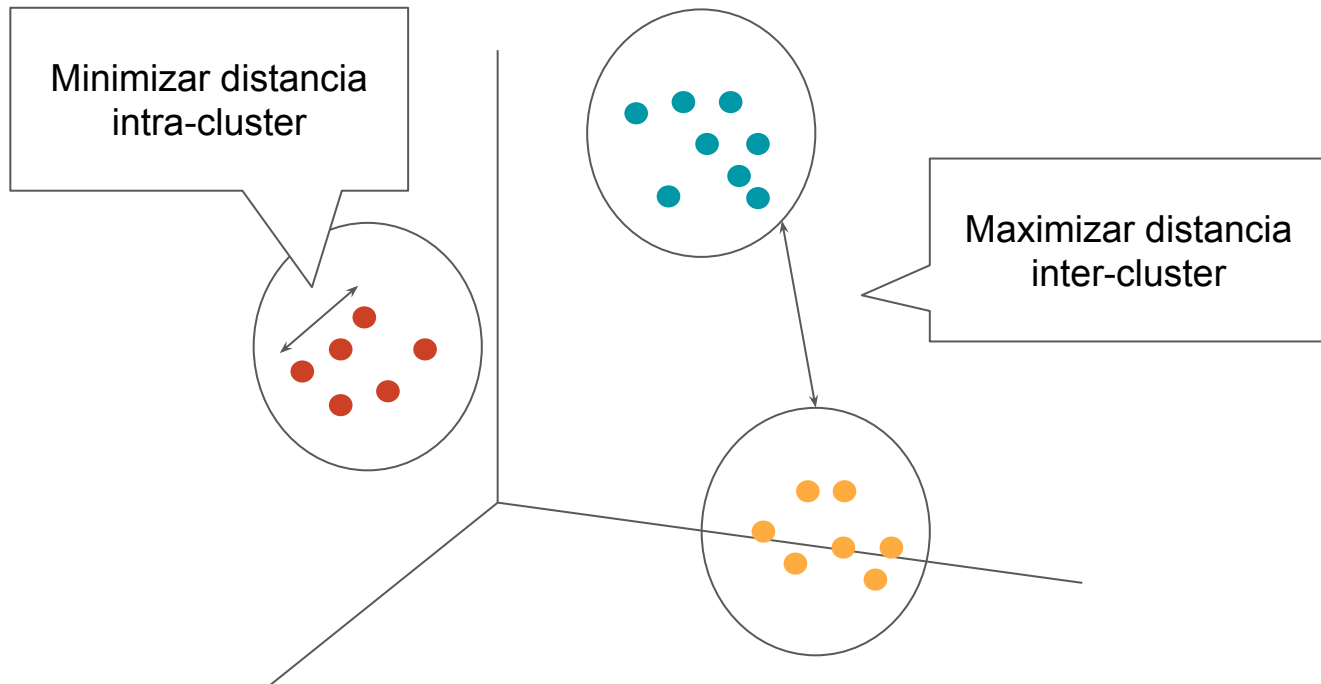
- Los objetos en un grupo sean similares (o relacionados) entre sí y,
- que sean diferentes (o no relacionados) a los objetos en otros grupos



Clustering



Clustering

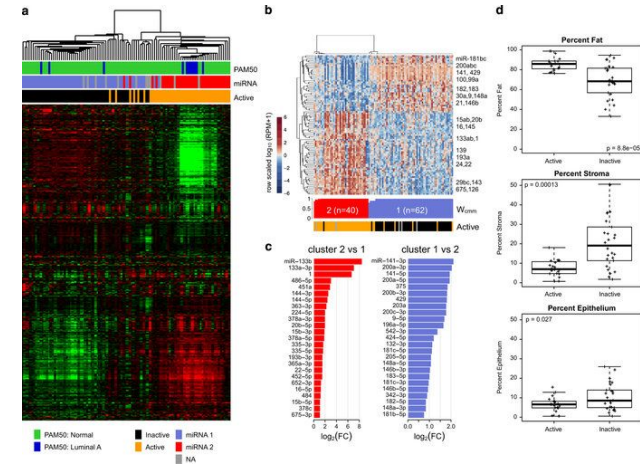


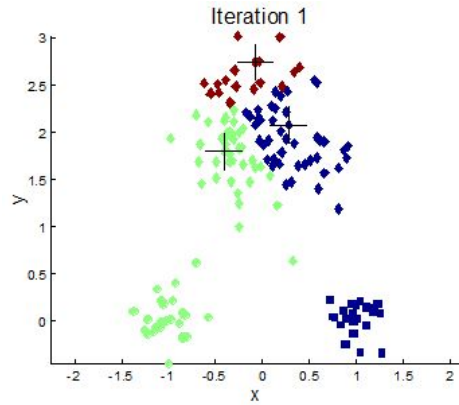
Ejemplos

- Agrupar imágenes para búsqueda.

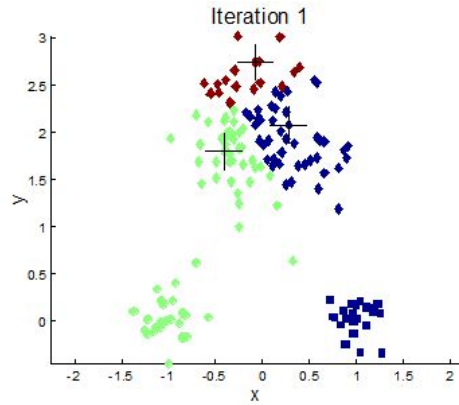


- Agrupar pacientes por condiciones médicas





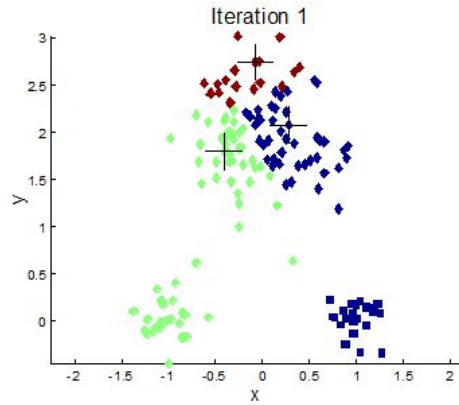
K-Means es uno de los métodos más conocidos



K-Means es uno de los métodos más conocidos



Partimos seleccionando K centroides
iniciales al azar.

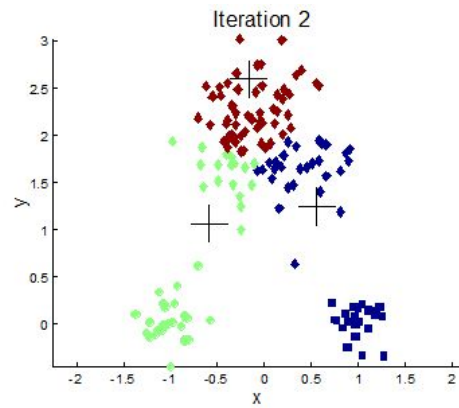
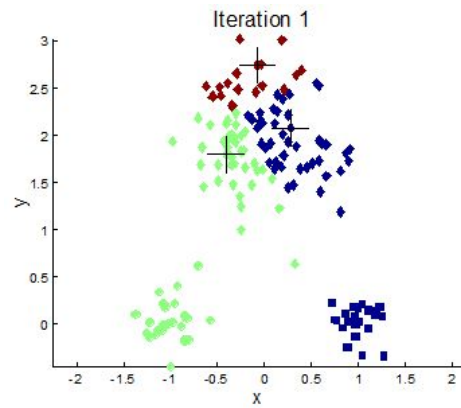


K-Means es uno de los métodos más conocidos

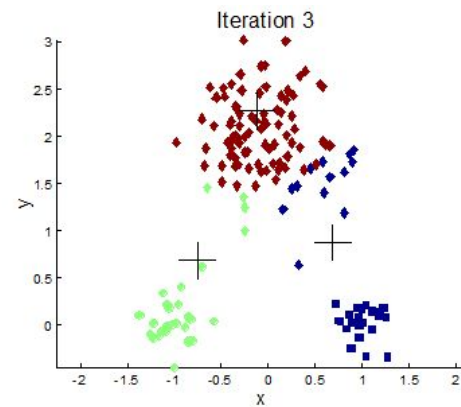
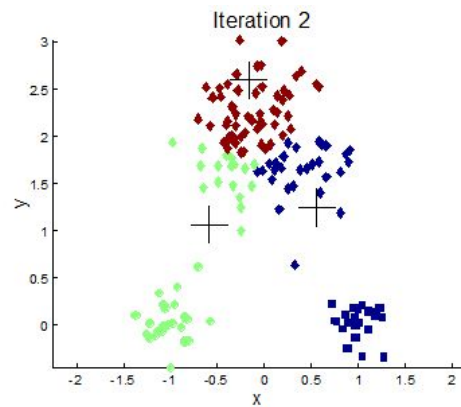
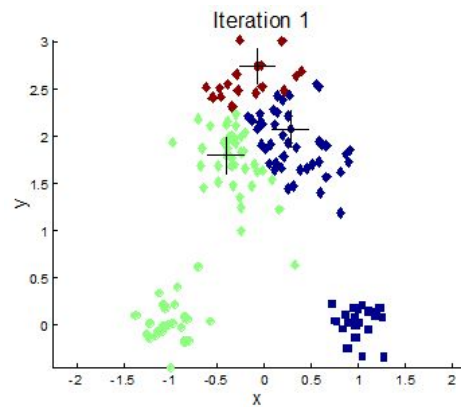


Partimos seleccionando K centroides
iniciales al azar.

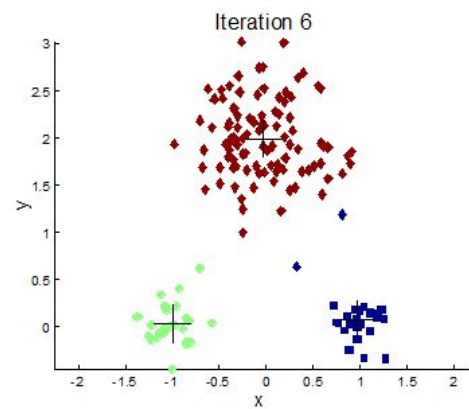
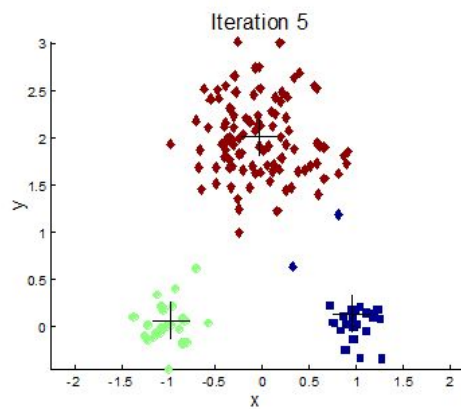
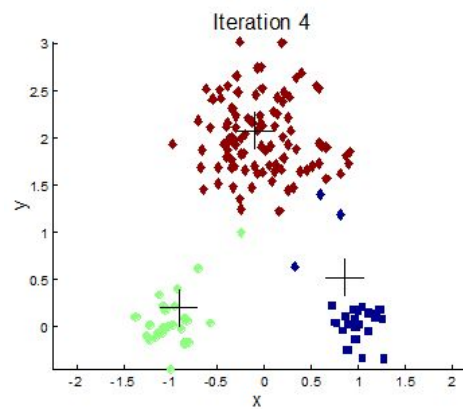
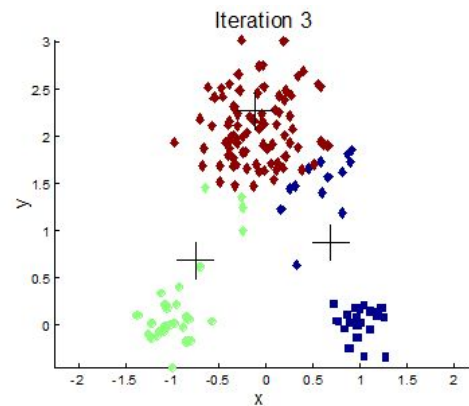
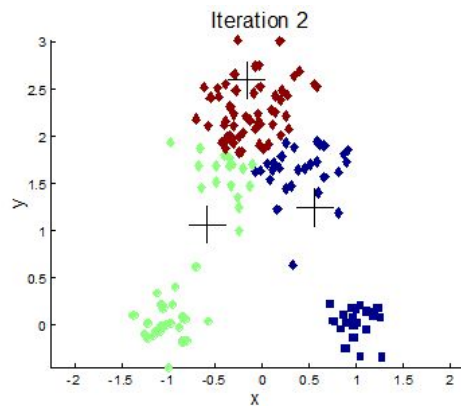
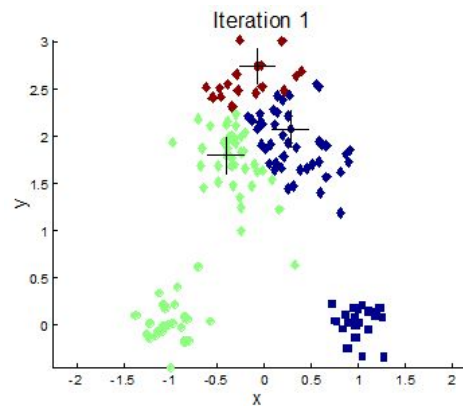
Asignamos cada elemento en el dataset al
centroide que esté más cerca.



Calculamos nuevos centroides



Iteramos



¿Son buenos nuestros clusters?

No hay respuesta absoluta -> depende de la aplicación

- Evitar encontrar patrones en el ruido
- Para comparar algoritmos de clustering diferentes
- Para comparar conjuntos de clusters diferentes
- Para comprar dos clusters

Evaluar clusterings

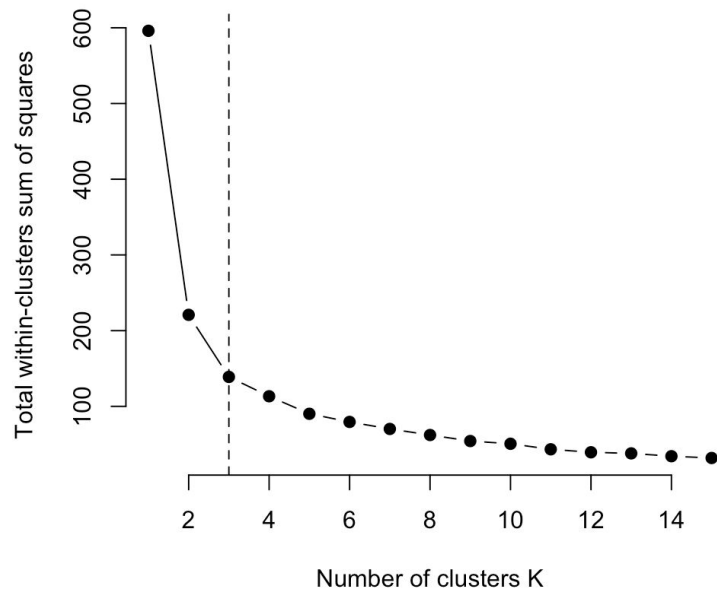
Suma error cuadrático: medida más común para evaluar clusters

- Por cada punto, error es la distancia al centroide del cluster
- x : punto en el cluster C_i , m_i : centroide C_i
- Dados 2 clusters se escoge el que tiene menos error

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} dist^2(m_i, x)$$

Método del codo

- Evaluar suma total de error cuadrático para distintos valores de K
- Buscar el “codo” que es el punto en que el error cuadrático baja de manera consistente



Resumen

Modelos de aprendizaje no supervisados: toman un conjunto de datos y crean una función que trate de darle sentido a datos de los que **no conocemos la estructura.**

Vimos el método K-Means.

-> Ir a jupyter notebook parte 3

Estructura general

Parte 1:

Preprocesamiento y Visualización

Parte 2:

Aprendizaje Supervisado

Parte 3:

Aprendizaje No Supervisado

¿Y redes neuronales artificiales?

¡Es un área muy grande! Algunos recursos:

- Curso en español del professor Jorge Pérez (UChile):
<https://github.com/dccuchile/CC6204>
- Libro profesor Alexandre Bergel (UChile):
<https://b-ok.cc/book/5679843/309d91> (está en el language Pharo pero está bien completo también)