



# 30.01 Semana 3

> Machine learning

# Introducción al machine learning

> J. Antonio García Ramírez  
jose.ramirez@cimat.mx



# Agenda

- Repaso (10)
- Selección de modelos
  - Grid search (5)
  - K-folds cv (15)
- Aprendizaje no supervisado
  - Cluster (5)
    - k-means (25)    `sys.sleep()`
    - Agrupación jerárquica (20)

# Agenda

- Aprendizaje no supervisado
  - Reducción de dimensión
    - PCA (35) `sys.sleep()`

# Precisión y tipos de error

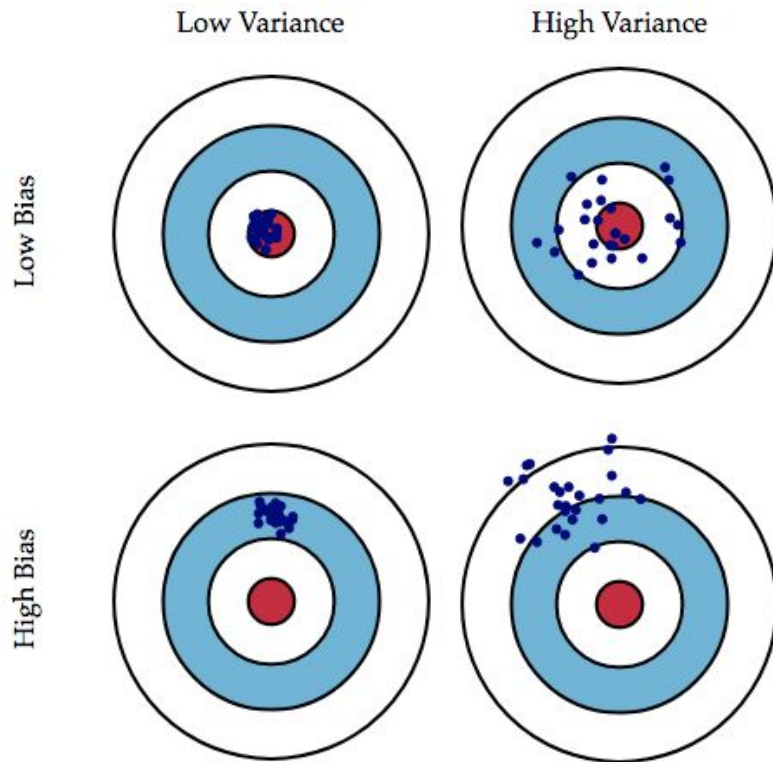
	Hypothesis true	Hypothesis false
Accept. hypothesis		Type II error
Reject hypothesis	Type I error	



# Tradeoff (1a parte que nos corresponde)

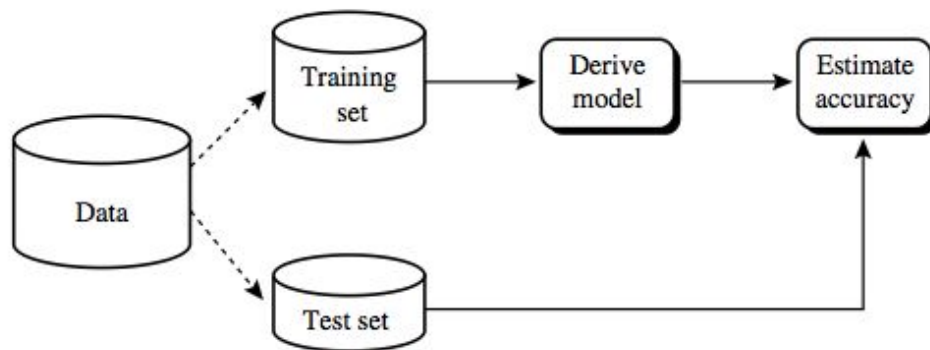
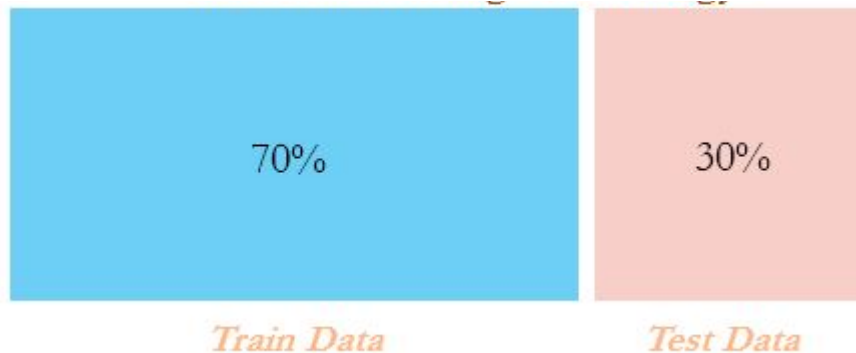
$$E \left( y_0 - \hat{f}(x_0) \right)^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon)$$

# Variance vs Bias tradeoff



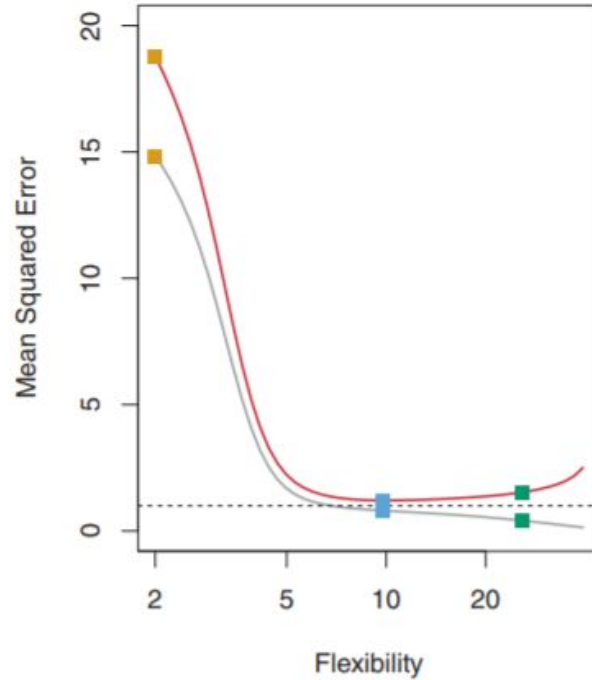


# Conjuntos train y test

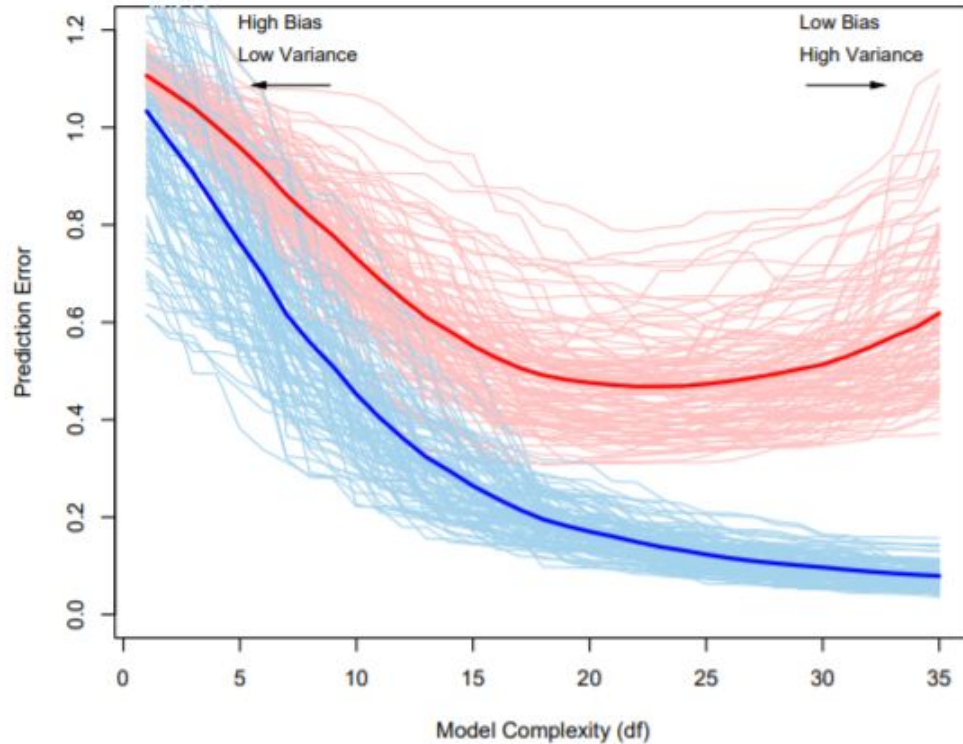


# Variance vs Bias tradeoff

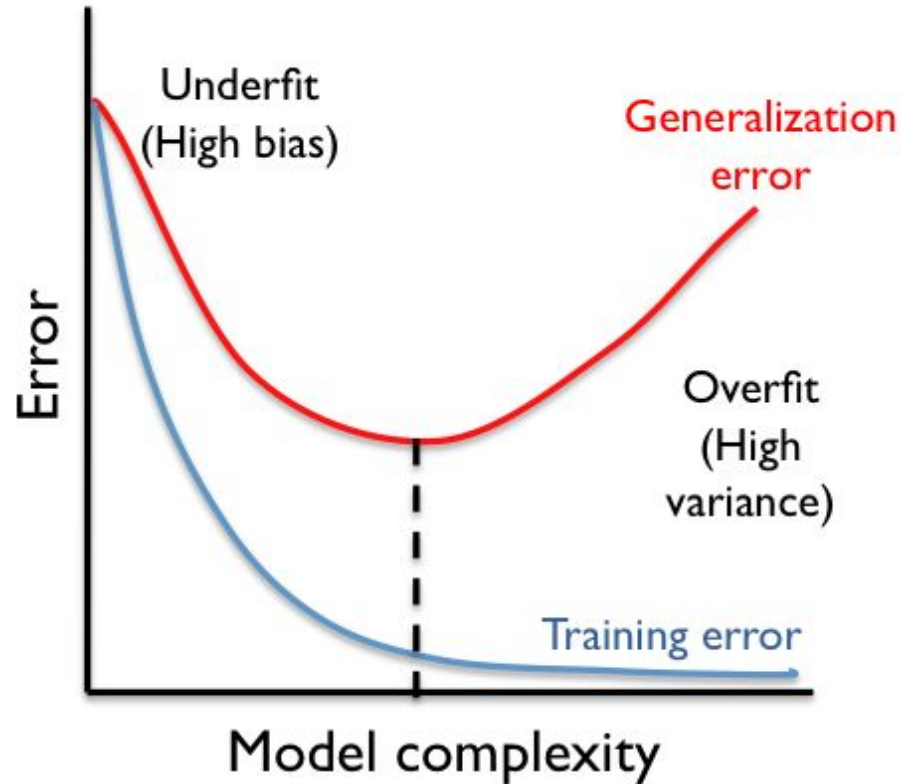
La carta a los reyes magos



# Variance vs Bias tradeoff



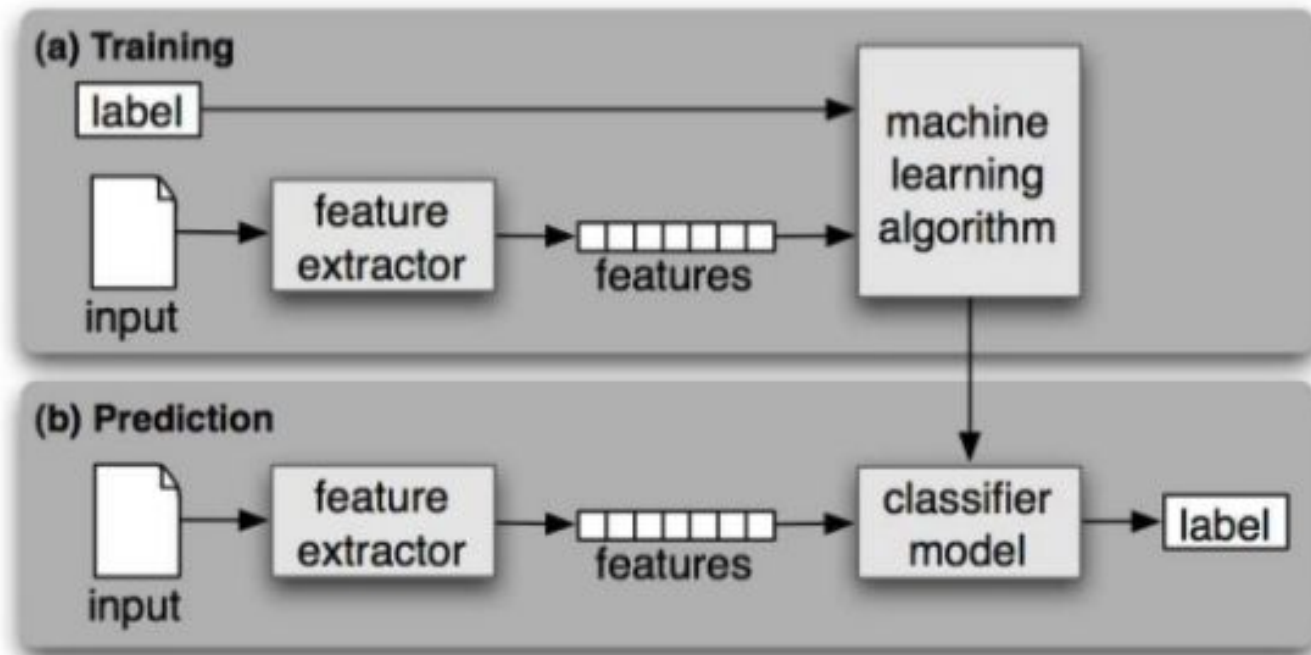
# Variance vs Bias tradeoff



# Tipos de variables

- Continuas (cuantitativas)
- Discretas (cualitativas)
  - Nominal
  - Ordinal
  - Categórica

# Flujo de trabajo



> Como vamos? manos a la obra:

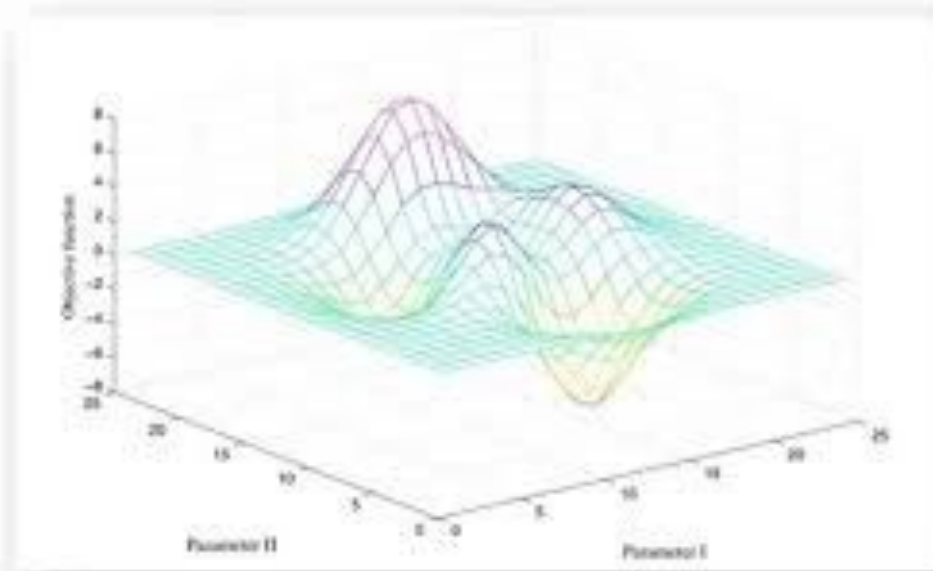
Plantea un proyecto de aprendizaje automático

# Selección de modelos

- Grid search
- K-folds cross validation



# Grid search



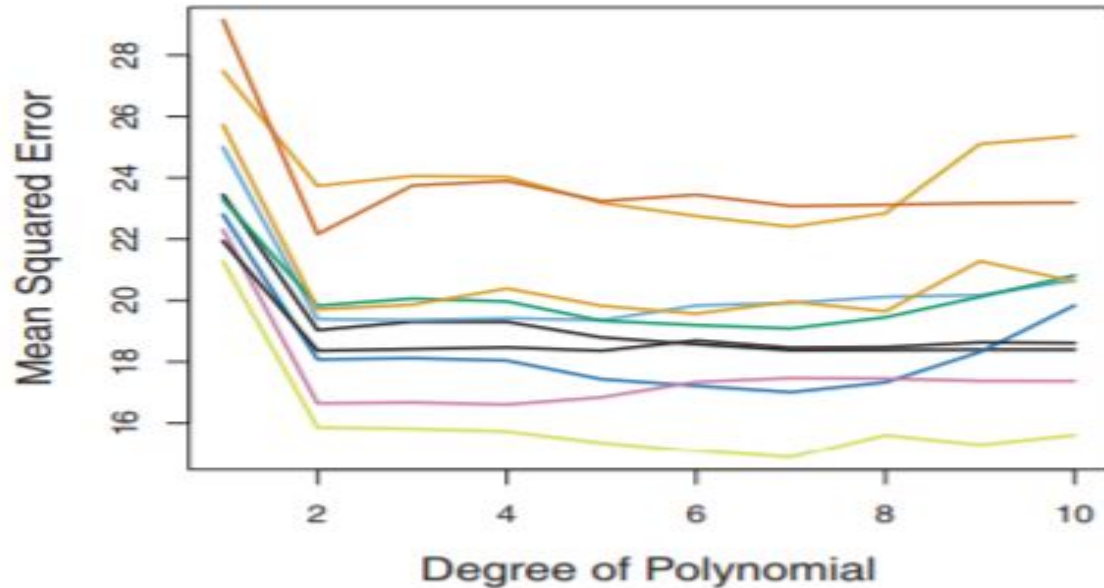
# K-folds cross validation

- Pocos datos



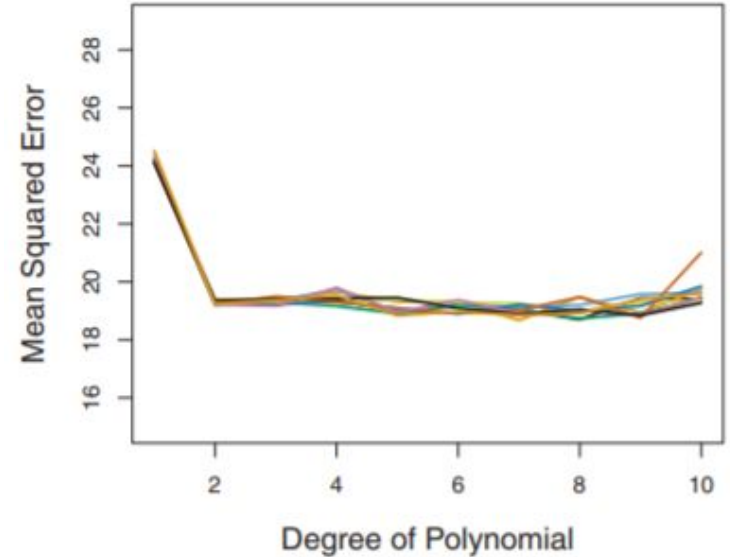
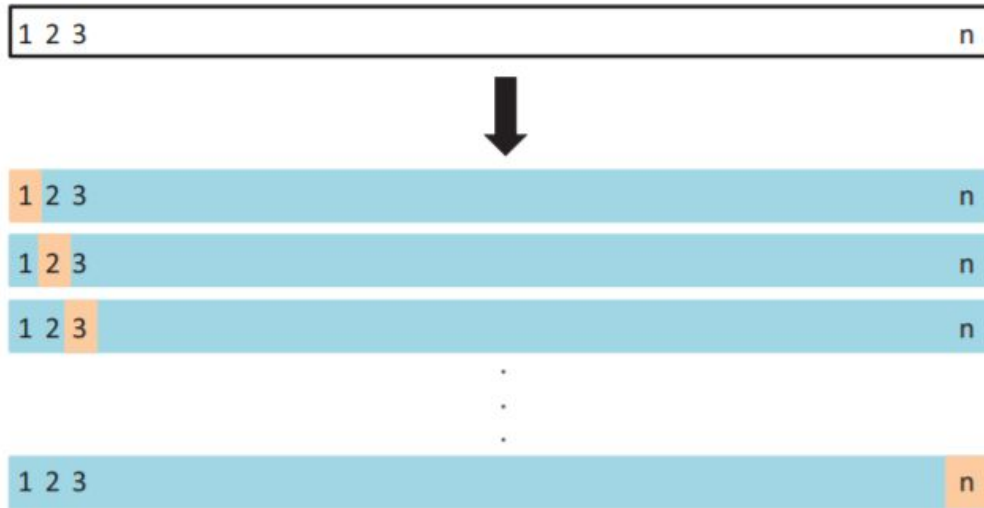
# K-folds cross validation

- Pocos datos, mucha varianza bajo sesgo



# Leave-one-out cv (LOOCV)

- Pocos datos, poca varianza y mucho sesgo



# Opinión

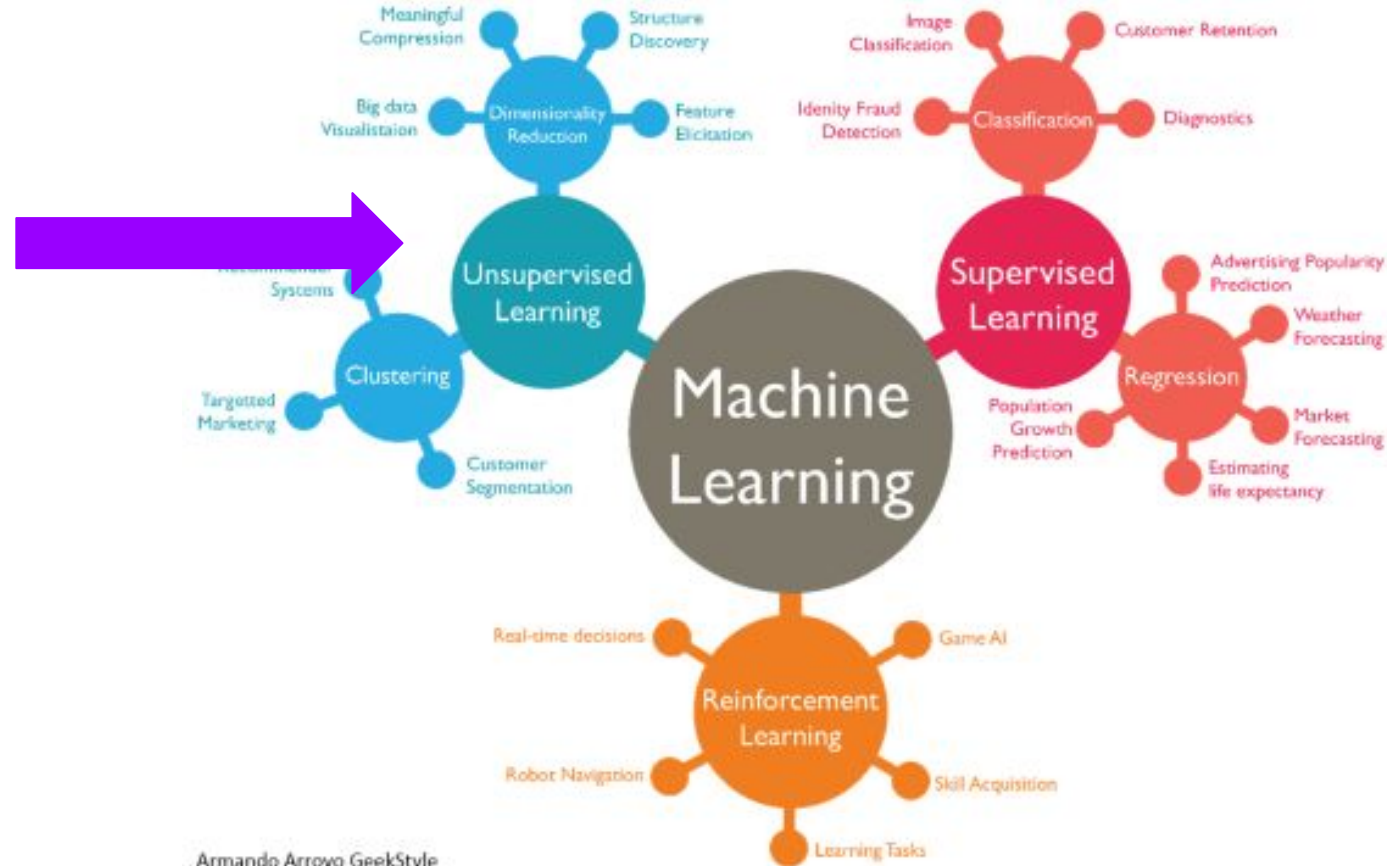
- “[...] no importa la pregunta, si estrujas lo suficiente los datos ellos te responderan [...]”

# Opinión

- “[...] no importa la pregunta, si estrujas lo suficiente los datos ellos te responderan [...]”

Sesgo

# El árbol, no supervisado



# Aprendizaje no supervisado

- Déf:
- Es una rama del ML que aprende de datos de pruebas que no se han etiquetado. En lugar de responder a la retroalimentación, **identifica puntos en común en los datos**



# Aprendizaje no supervisado

- Una aplicación central del aprendizaje no supervisado se encuentra en el campo de la estimación de la densidad (kernels)



# Clustering

Se refiere a un conjunto (muy amplio) de técnicas para encontrar subgrupos o agrupación clusters, en un conjunto de datos.

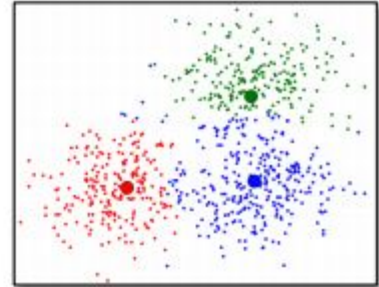
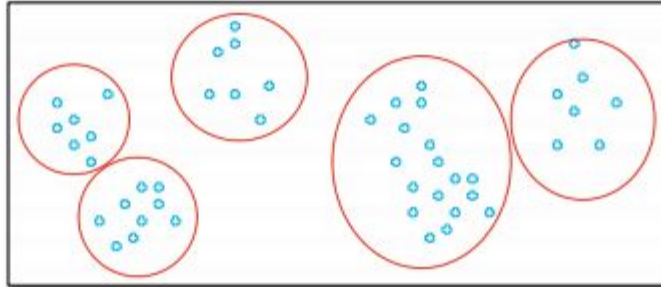
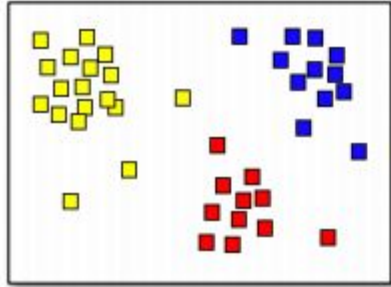
# Clustering

Dividir los datos en grupos distintos:

Observaciones **dentro** de cada grupo sea **muy similar** entre sí, mientras que las observaciones **en diferentes grupos** son **muy diferentes** entre sí

Qué significa que dos o más observaciones sean similares o diferentes ? (específico del dominio)

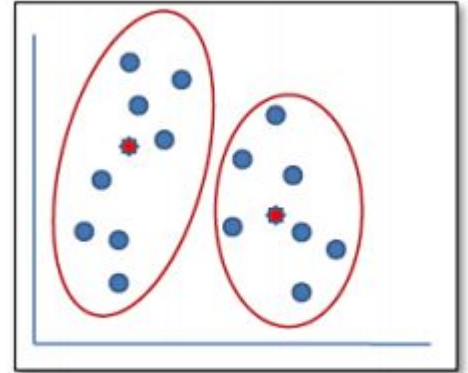
# Clustering



# K-means

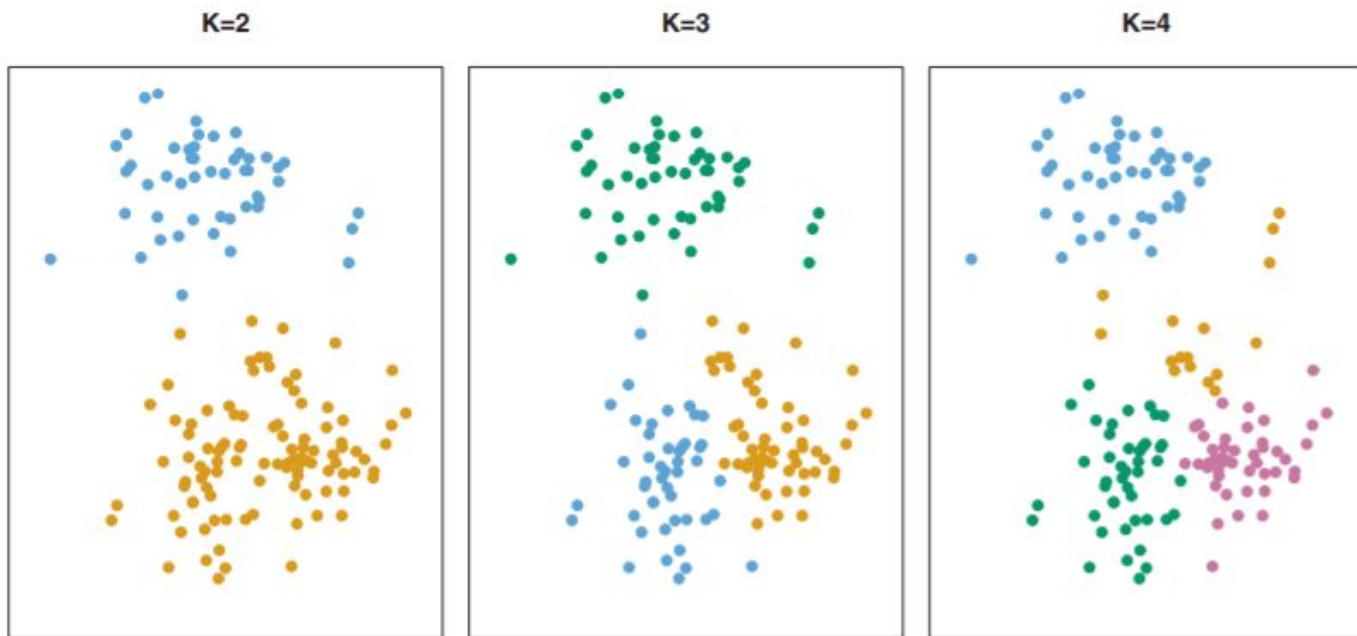
Es un enfoque **simple** y elegante para la partición de un conjunto de datos en  $K$  grupos distintos, **no superpuestos**.

Debemos **especificar** el número deseado de agrupaciones  $K$ , entonces el algoritmo asignará cada observación a exactamente uno de los  $K$  grupos



# K-means

Dado el número de puntos, encontrar los centroides que minimizan la **distancia** con relación a las observaciones



# K-means

$$\underset{C_1, \dots, C_K}{\text{minimize}} \left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right\}$$

## Algoritmo 1: k-means

- 1) Asigne aleatoriamente un número, de 1 a  $K$ , a cada una de las observaciones. Estos sirven como asignaciones de grupo iniciales para las observaciones

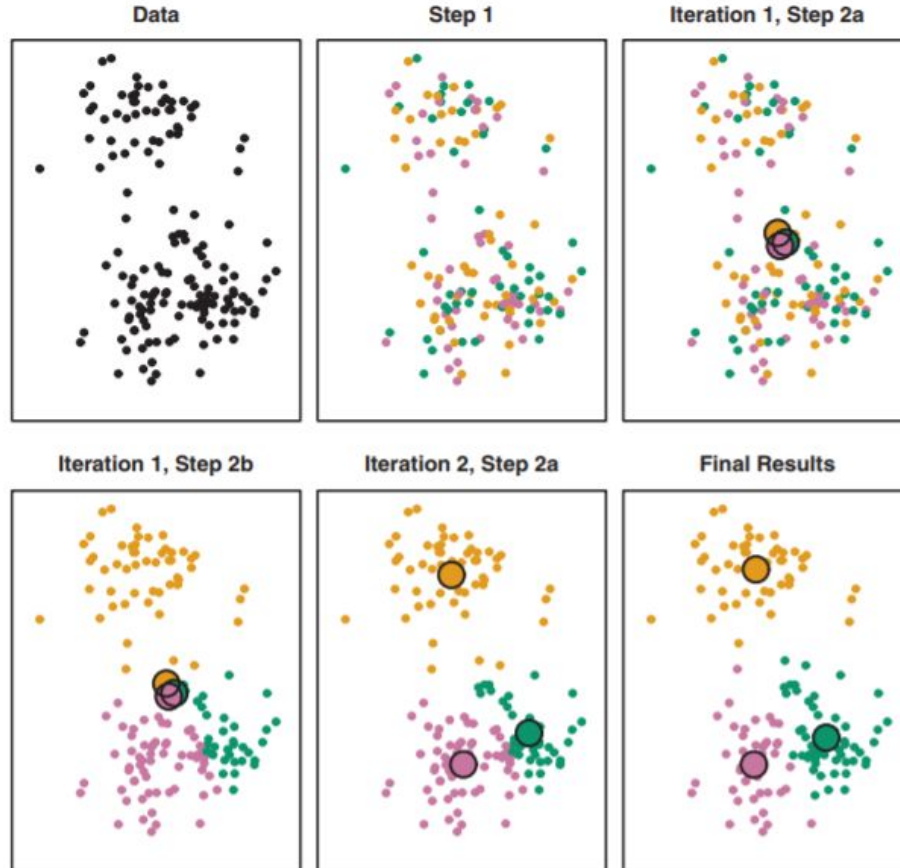
# K-means

## Algoritmo 1: k-means

- 2) Iterar hasta que las asignaciones de clúster dejen de cambiar:
  - a) Para cada uno de los  $K$  clusters, calcule el centroide del cluster. El centroide de clúster  $k$ -ésimo es el vector de las medias de la característica  $p$  para las observaciones del  $k$ -ésimo grupo
  - b) Asigne cada observación al grupo cuyo centroide esté más cerca (utilizando la distancia euclidiana).



# K-means



Relación con la  
teselación de  
Voronoi

[https://upload.wikimedia.org/wikipedia/commons/e/ea/K-means\\_convergence.gif](https://upload.wikimedia.org/wikipedia/commons/e/ea/K-means_convergence.gif)

> Ejemplo interactivo

k-means

*[https://foufoo.shinyapps.io/mp\\_kmeans\\_cpp/](https://foufoo.shinyapps.io/mp_kmeans_cpp/)*

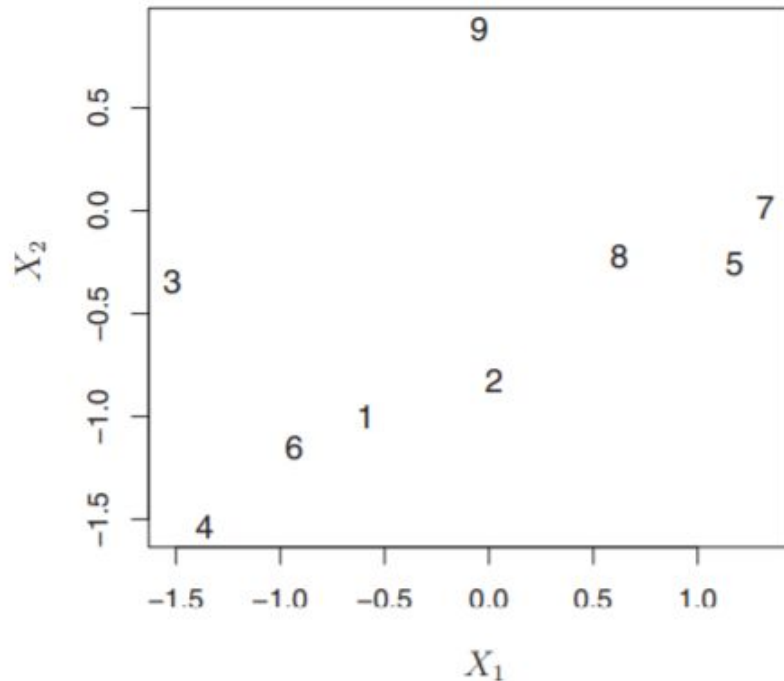
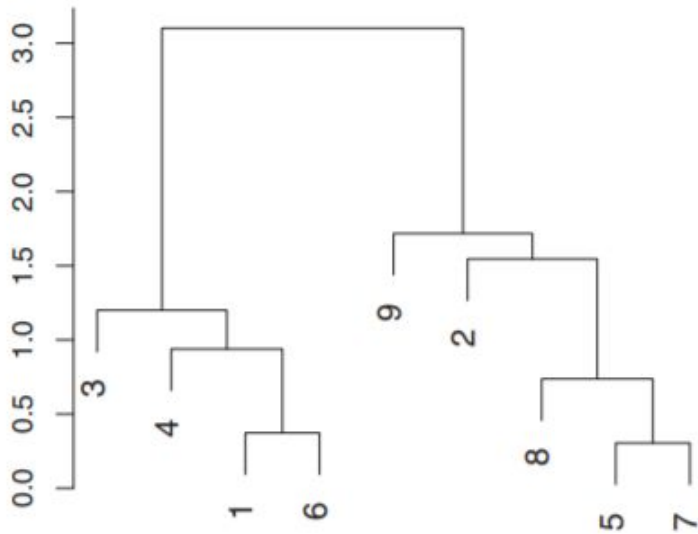
# Agrupación jerárquica

Una desventaja de K-means es que pre-especifique el número  $K$ .

El clustering jerárquico no requiere que nos comprometemos a la elección de  $K$ . Se traduce en una representación basada en árboles de las observaciones, **dendrogramas**.

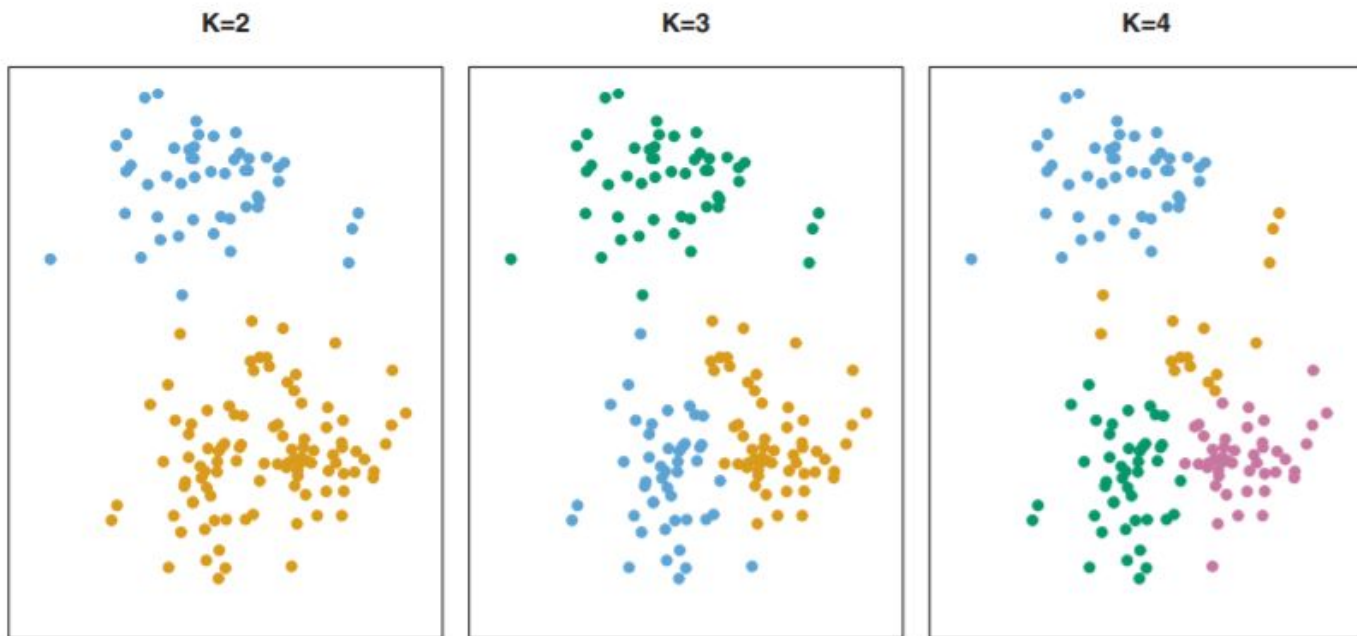
Describimos el agrupamiento de abajo hacia arriba o aglomerado.

# Agrupación jerárquica



# K-means

Dado el número de puntos, encontrar los centroides que minimizan la **distancia** con relación a las observaciones



# Agrupación jerárquica

## Algoritmo 2: Agrupación jerárquica

- 1) Comience con  $n$  observaciones y una medida (como la distancia euclidiana) de todas las disimilitudes entre pares. Tratar cada observación como su propio cluster.

# Agrupación jerárquica

## Algoritmo 1: Agrupación jerárquica

2) Para  $i = n, n - 1, n - 2, \dots, 2$

- a) Explorar todas las disimilitudes entre clústeres por pares entre los  $i$  grupos e identificar el par de agrupaciones que son menos disímiles (lo más parecido). Fusiona estos dos grupos. La disimilitud entre estos dos grupos indica la altura en el dendrograma en el que debe colocarse la fusión.
- b) Calcule las nuevas disimilitudes intercluster entre pares de los  $i - 1$  restantes

# Agrupación jerárquica

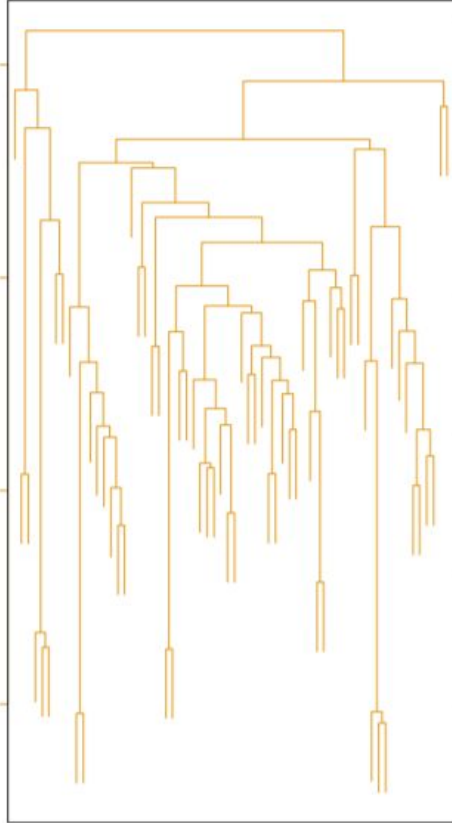
Formas usuales de medir disimilaridades:

- Complete
- Single
- Average
- Centroid

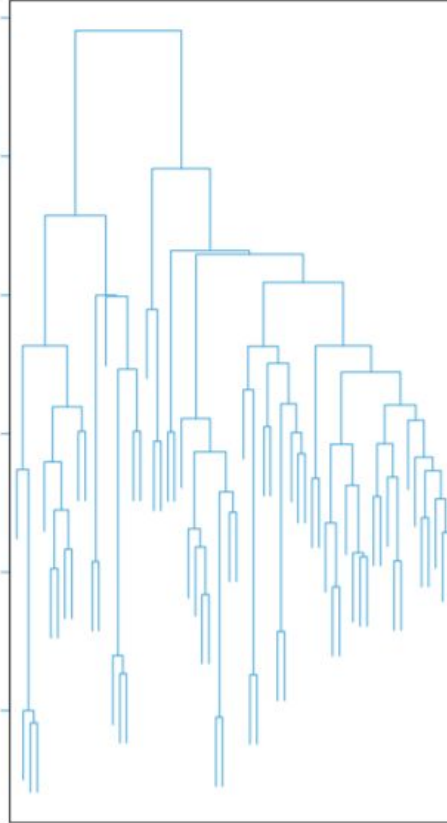


# Agrupación jerárquica

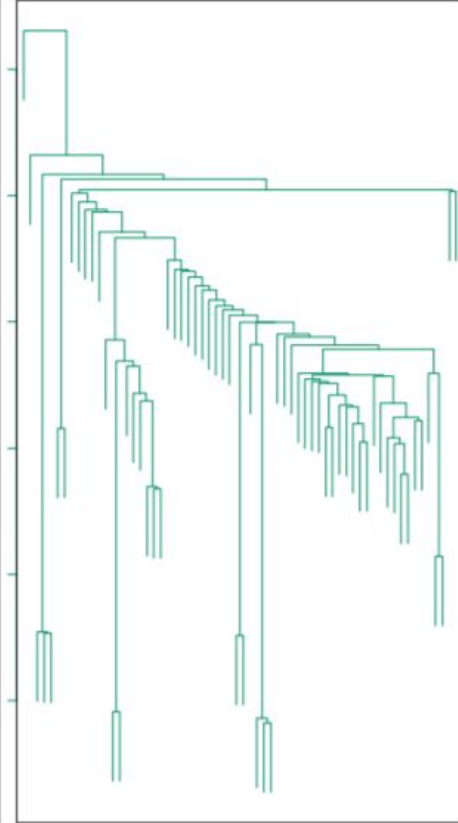
Average Linkage



Complete Linkage



Single Linkage



# Clusterización

Para qué sirve ?

- Segmentar muestra, mercado, clientes, transacciones, ...
  - Mayor estabilidad de parámetros
    - Menos sesgo y varianza
  - Un modelo diferente en cada grupo

Ejemplos: AGH, AGGC, ...

# Reducción de dimensión

Qué pasa cuando tenemos muchas features ?

- La maldición de la dimensionalidad
  - Parámetros inestables
  - Muestreo insuficiente
    - Ejemplo:  $\frac{1}{2}$  millón de obs. de dimensión 90 equivalen a 2 obs. de una sola dimensión



# Principal Components Analysis (PCA)

¿Para qué los componentes principales?

Visualizar observaciones con mediciones en un conjunto de  $p$  variables,  $X_1, X_2, \dots, X_p$  como parte de un análisis exploratorio de datos. Usando diagramas de dispersión bidimensionales de los datos, cada uno de los cuales contiene las mediciones de dos en dos hay  $\binom{n}{2} = \frac{n(n-1)}{2}$

# Principal Components Analysis (PCA)

¿Para qué los componentes principales?

Por ejemplo con  $p = 10$  hay 45 pares! Si  $p$  es grande, entonces ciertamente no será posible mirar a todos pares de manera significativa.

Podemos obtener una representación bidimensional de los datos que capturan la mayor parte de la información. En un espacio de baja dimensión.

# Principal Components Analysis (PCA)

¿Para qué los componentes principales?

- Menos variables, menor tiempo de cómputo :D
- Se pierde interpretabilidad

# Principal Components Analysis (PCA)

¿Qué son los componentes principales?

La primer componente principal es la comb.  
lineal

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p$$

Que maximiza la varianza muestral, la segunda es ortogonal a la primera y su vez también maximiza el resto de la varianza muestral, la tercera ...

# Principal Components Analysis (PCA)

$$\underset{\phi_{11}, \dots, \phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^p \phi_{j1}^2 = 1.$$

¿

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p$$

Qué son los componentes principales?

La primer componente principal es la comb.  
lineal



# Principal Components Analysis (PCA)

$$\underset{\phi_{11}, \dots, \phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^p \phi_{j1}^2 = 1.$$

Más visual

<https://i.stack.imgur.com/Q7HIP.gif>

Son únicas ?

# Principal Components Analysis (PCA)

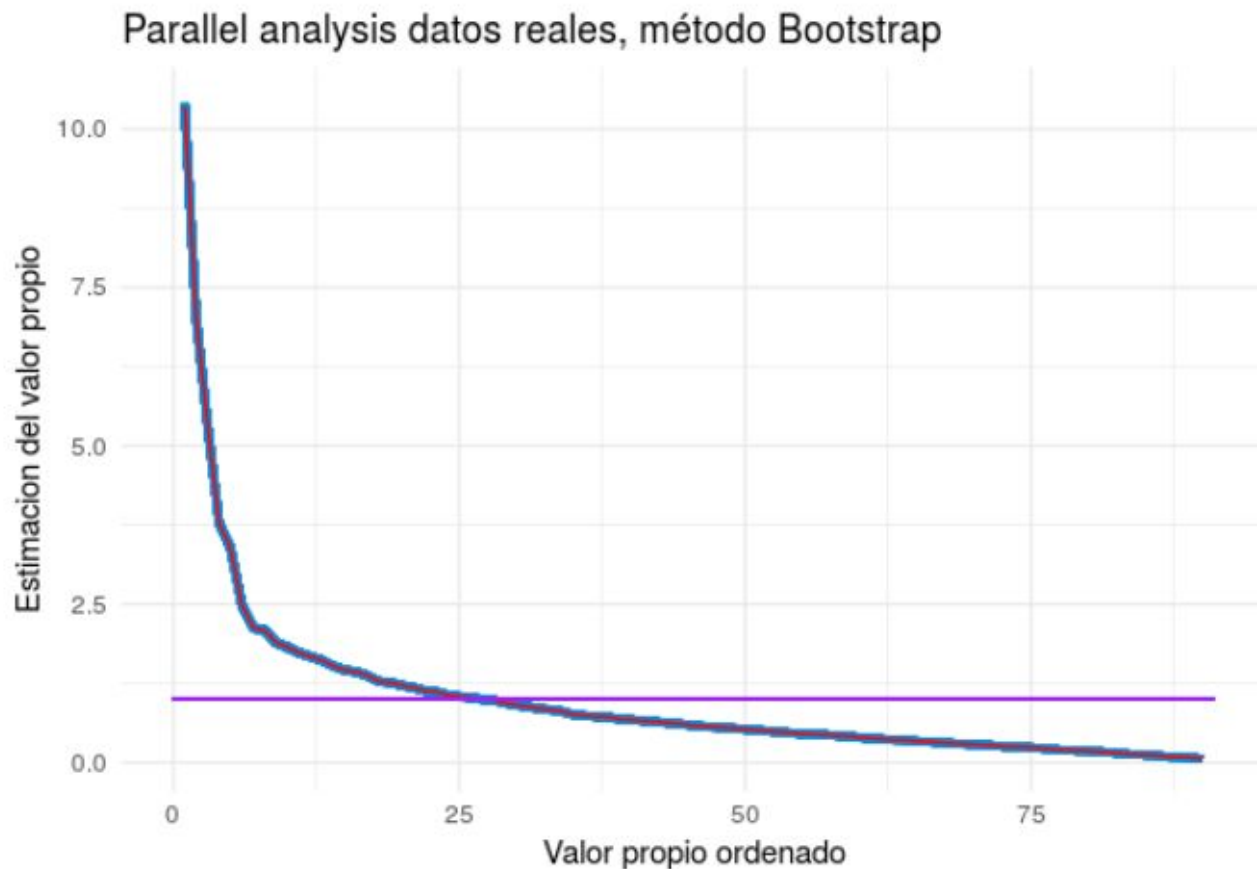
## Algoritmo 3: Cálculo de PC

- 1) Determine la matriz de covarianzas de su muestra  $p$  dimensional  $S$
- 2) Obtenga una descomposición de valores y vectores propios de  $S$ , los vectores propios forman las PC
- 3) Los vectores propios determinan la varianza explicada

¿ Cuantos componentes utilizar?

(PCA)

Teorema:  
varianza total



¿ Cuantos componentes utilizar?

(PCA)

Teorema: Los necesarios

<http://jkunst.com/flexdashboard-highcharter-examples/pokemon/>

## > **Actividad**

Sigue el Jupyter Notebook de clustering.

Clusterea las ciudades por distancia, temperatura, aumentando las K

> Go further

- Bootstrap
- Hyperparameter optimization
- Kernels
- K-medoids Clustering
- ISOMAP