Pontificia Universidad JAVERIANA Bogotá

VIGILADA MINEDUCACIÓN

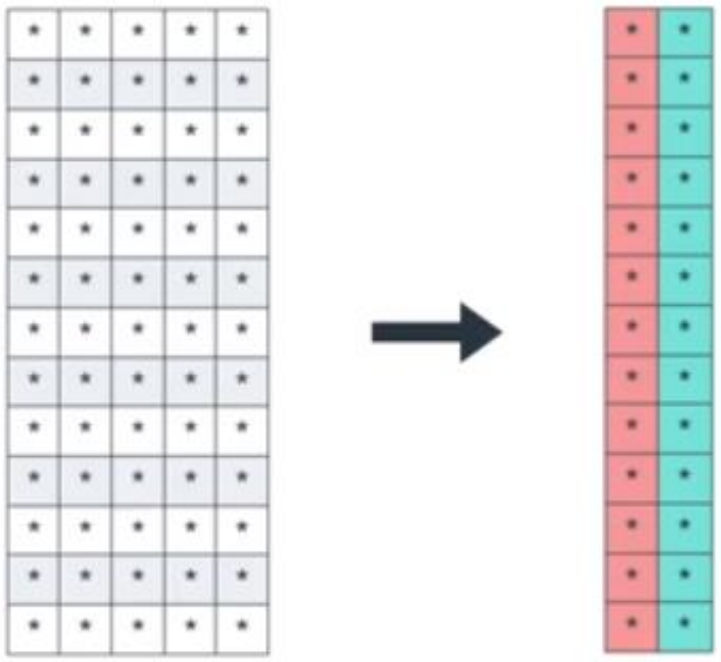
Reducción de la dimensionalidad
ANÁLISIS DE COMPONENTES PRINCIPALES

Pontificia Universidad JAVERIANA
Colombia

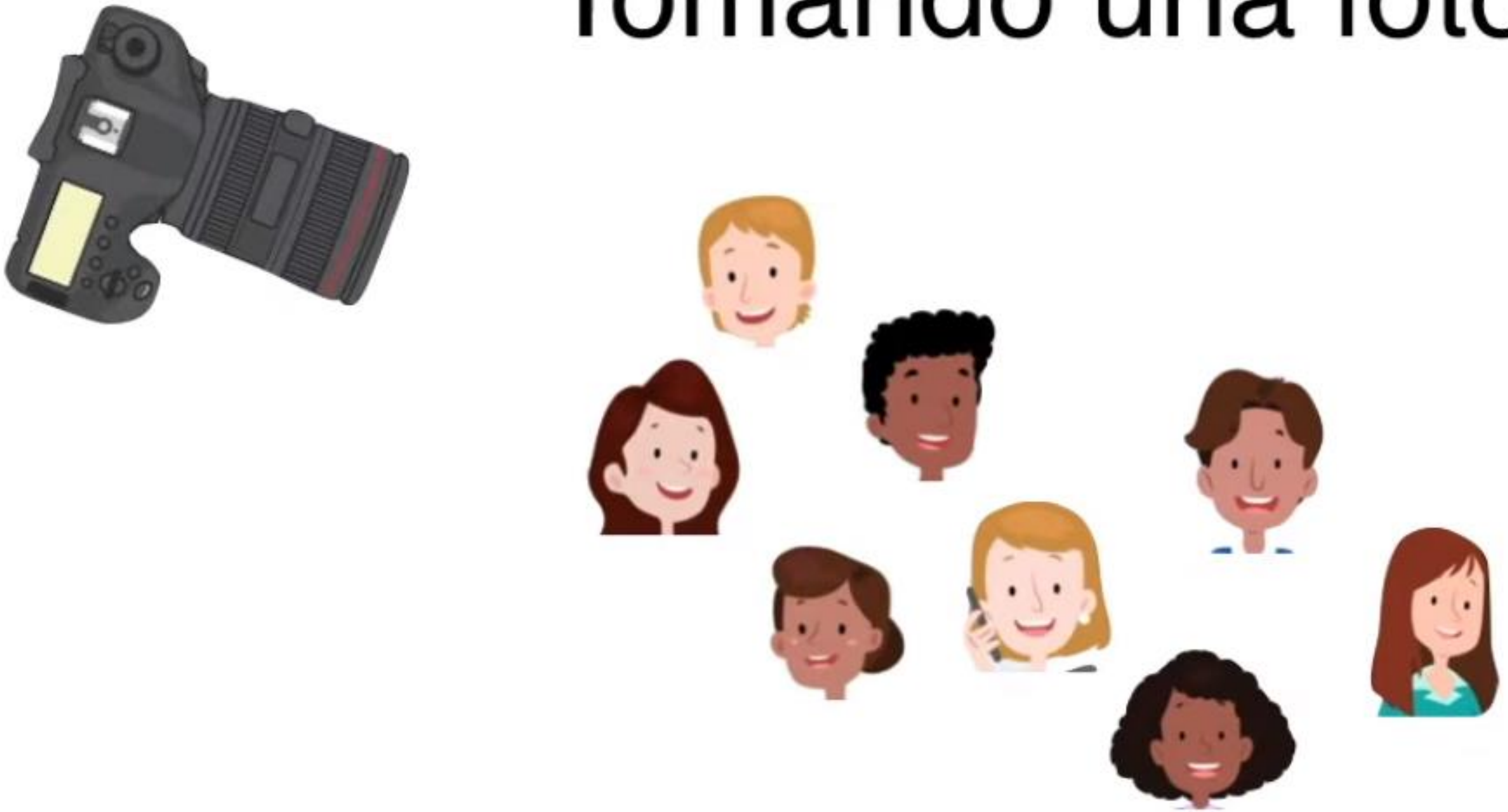
La Maldición de la Dimensionalidad

Muchas variables implican un crecimiento exponencial de los datos necesarios para obtener resultados significativos porque la información queda dispersa en el hiperespacio.

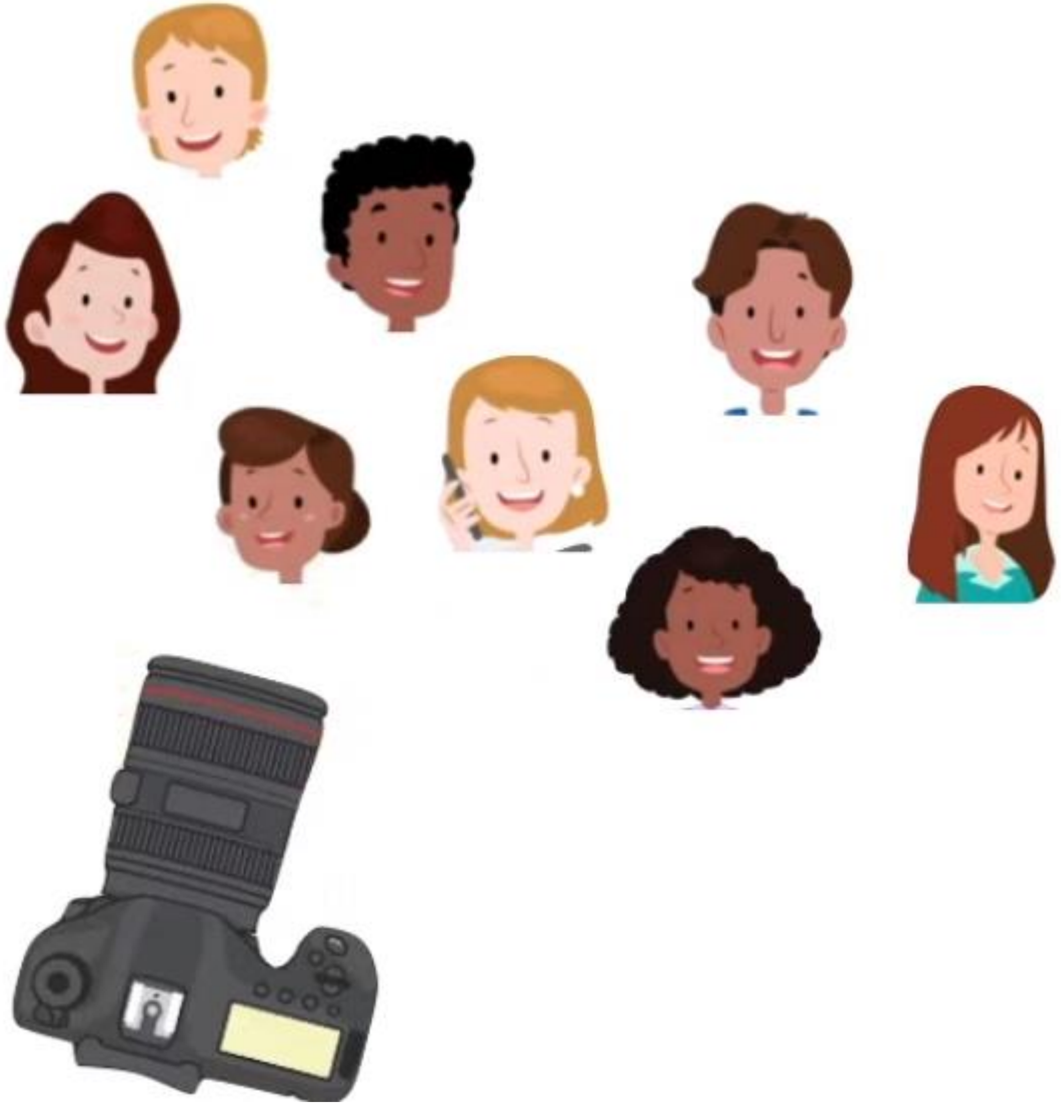
Reduccion de dimensionalidad



Tomando una foto

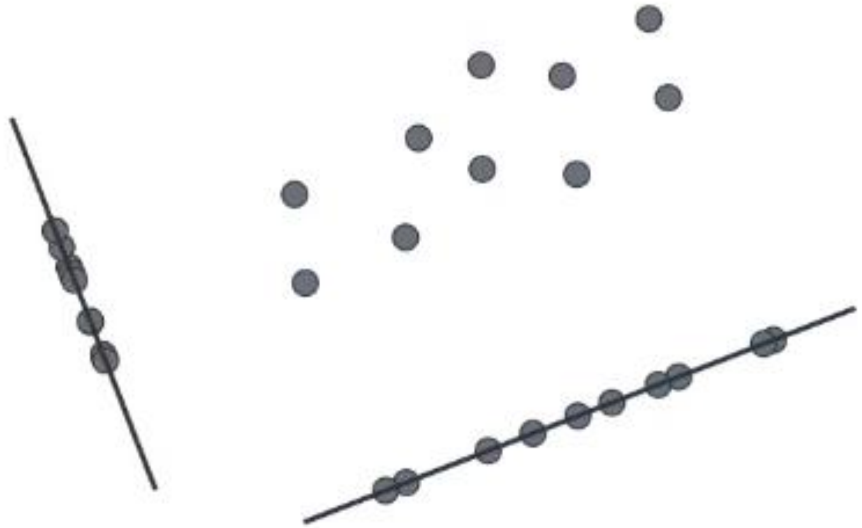


Tomando una foto

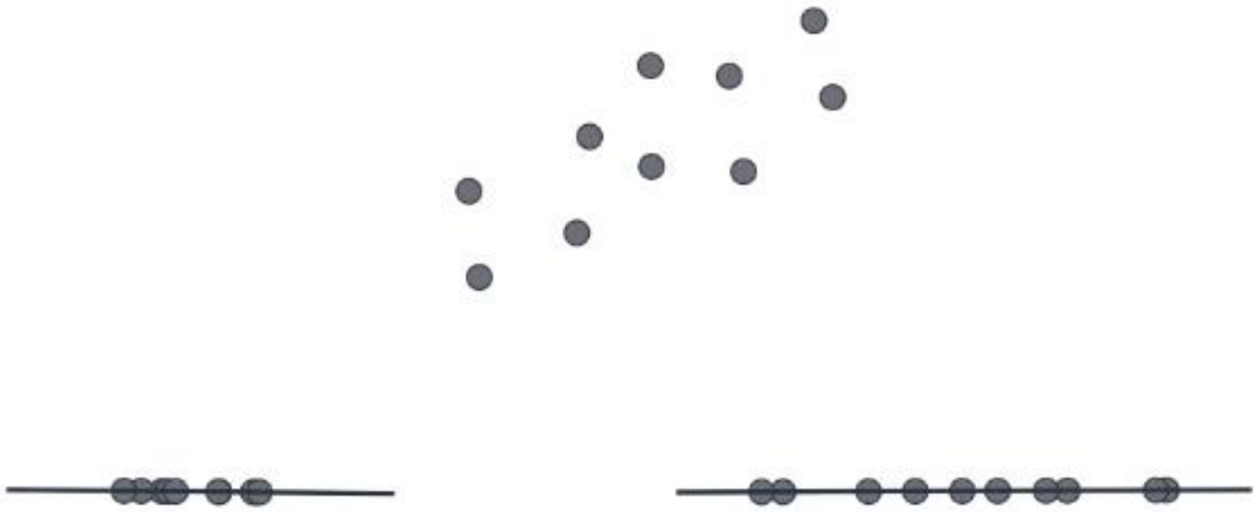


Reduccion de dimensionalidad

Reduccion de dimensionalidad



Reduccion de dimensionalidad



Datos de casas

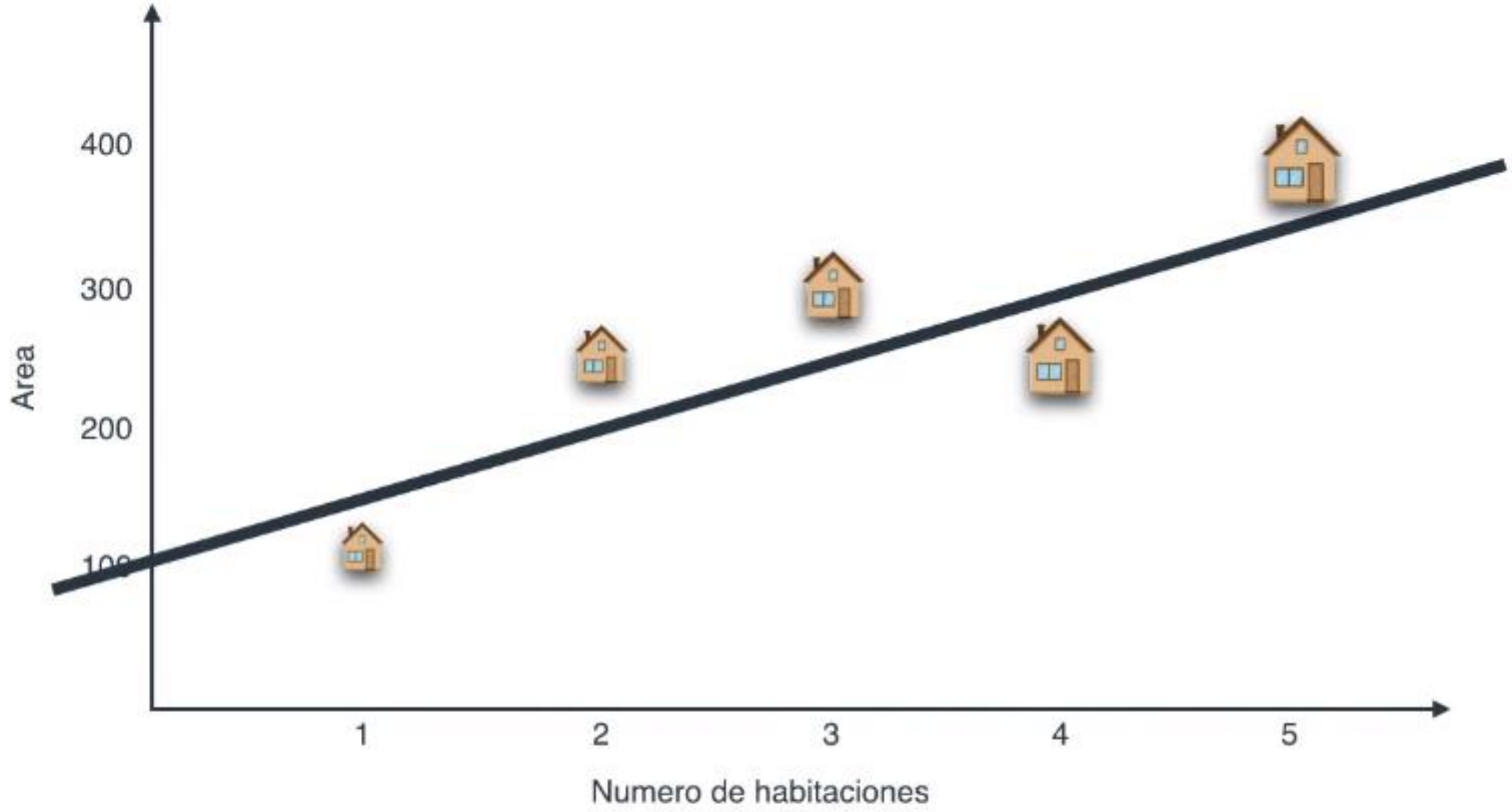
Area
Numero de habitaciones
Numero de baños
Escuelas cercanas
Crimen en el area

Datos de casas

Area
Numero de habitaciones → Tamaño
Numero de baños

Escuelas cercanas
Crimen en el area → Ubicación


400
300
200
100
Area
1
2
3
4
5
Numero de habitaciones


400
300
200
100
Area
1
2
3
4
5
Numero de habitaciones


Tamaño

2 dimensiones

area
numero de habitaciones

1 dimension

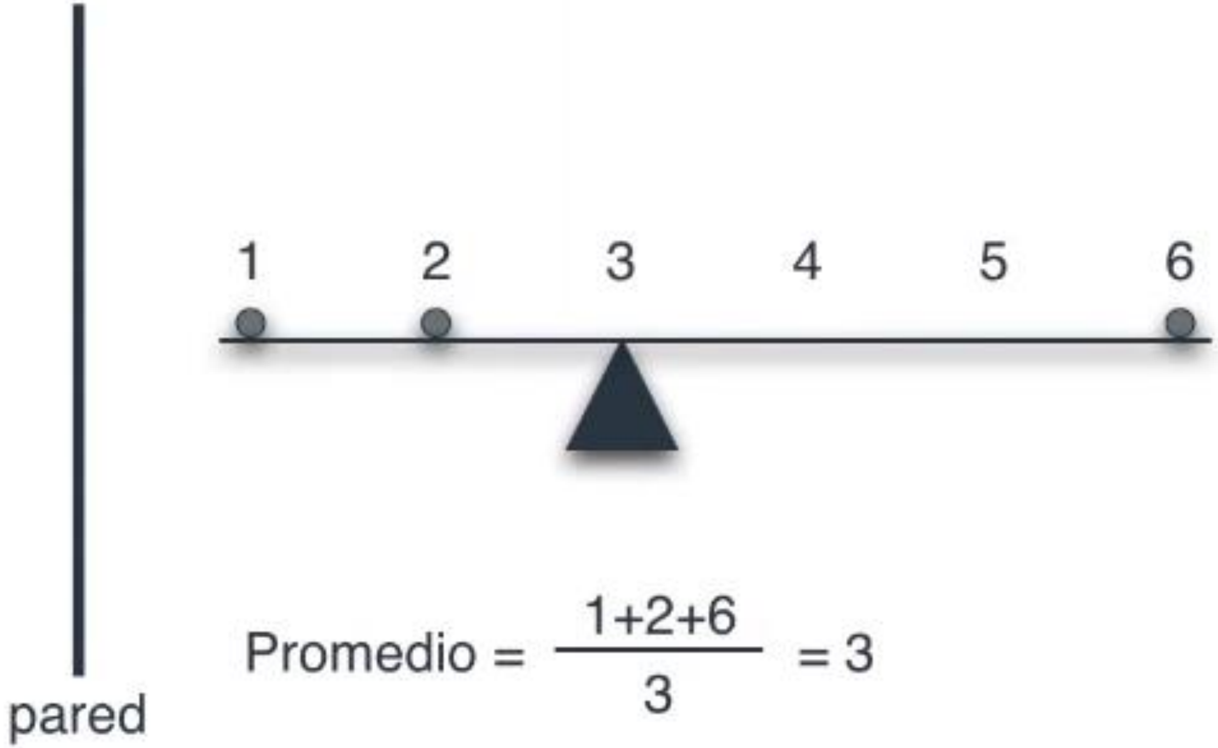
tamaño

Datos de casas

5 dimensiones	**2 dimensions**
Area Numero de habitaciones Numero de baños	Tamaño
Escuelas cercanas crimen en el area	Ubicación

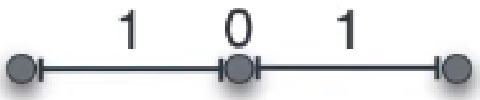
Promedio, varianza, covarianza

Promedio

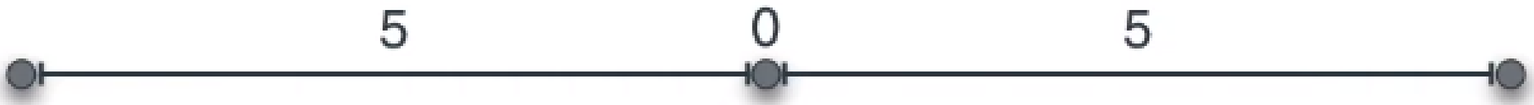


$$\text{Promedio} = \frac{1+2+6}{3} = 3$$

Varianza



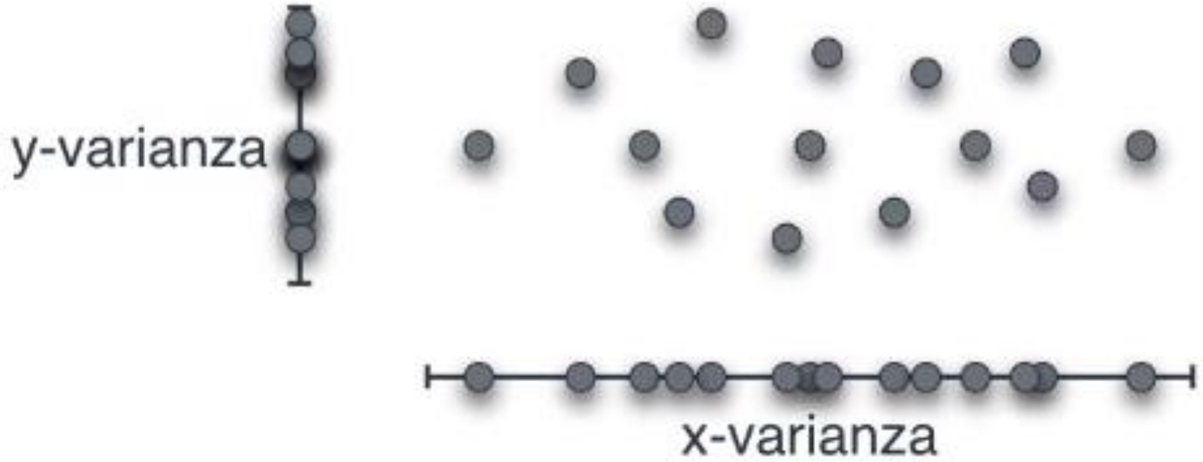
$$\text{Varianza} = \frac{1^2+0^2+1^2}{3} = 2/3$$



$$\text{Varianza} = \frac{5^2+0^2+5^2}{3} = 50/3$$

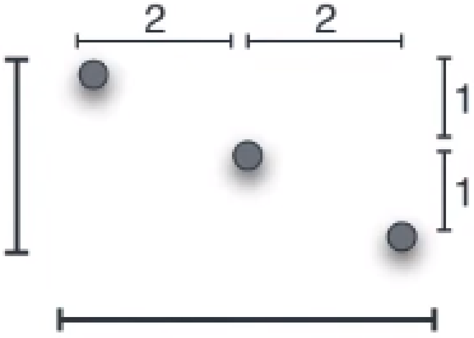
Varianza?

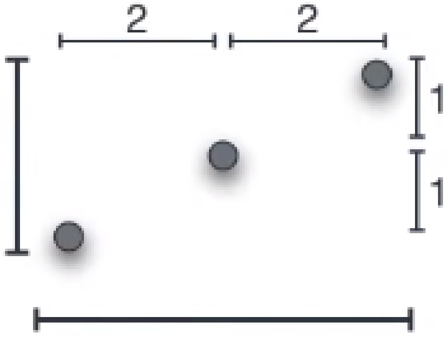
Varianza?



Varianza?

Varianza?

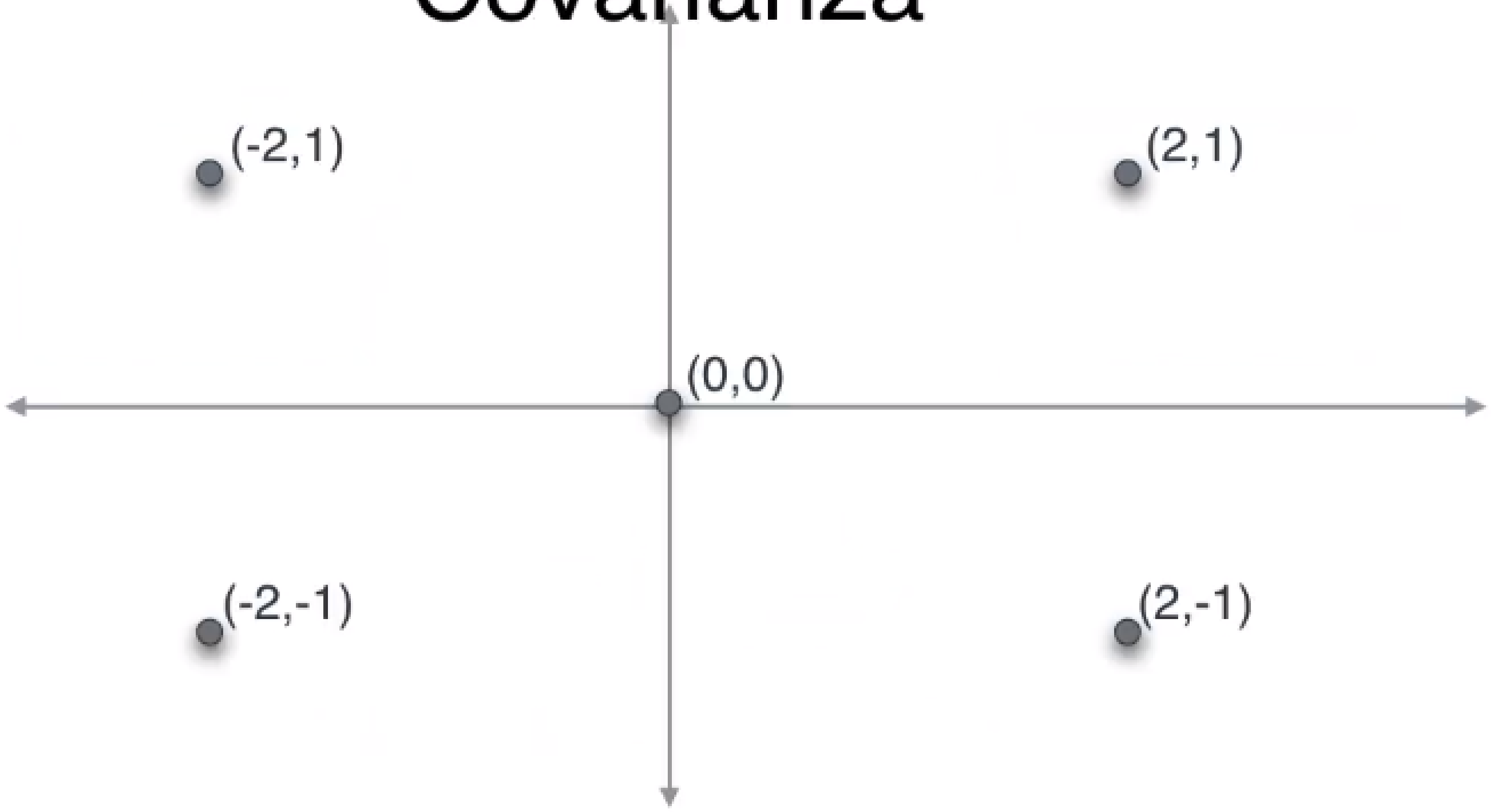


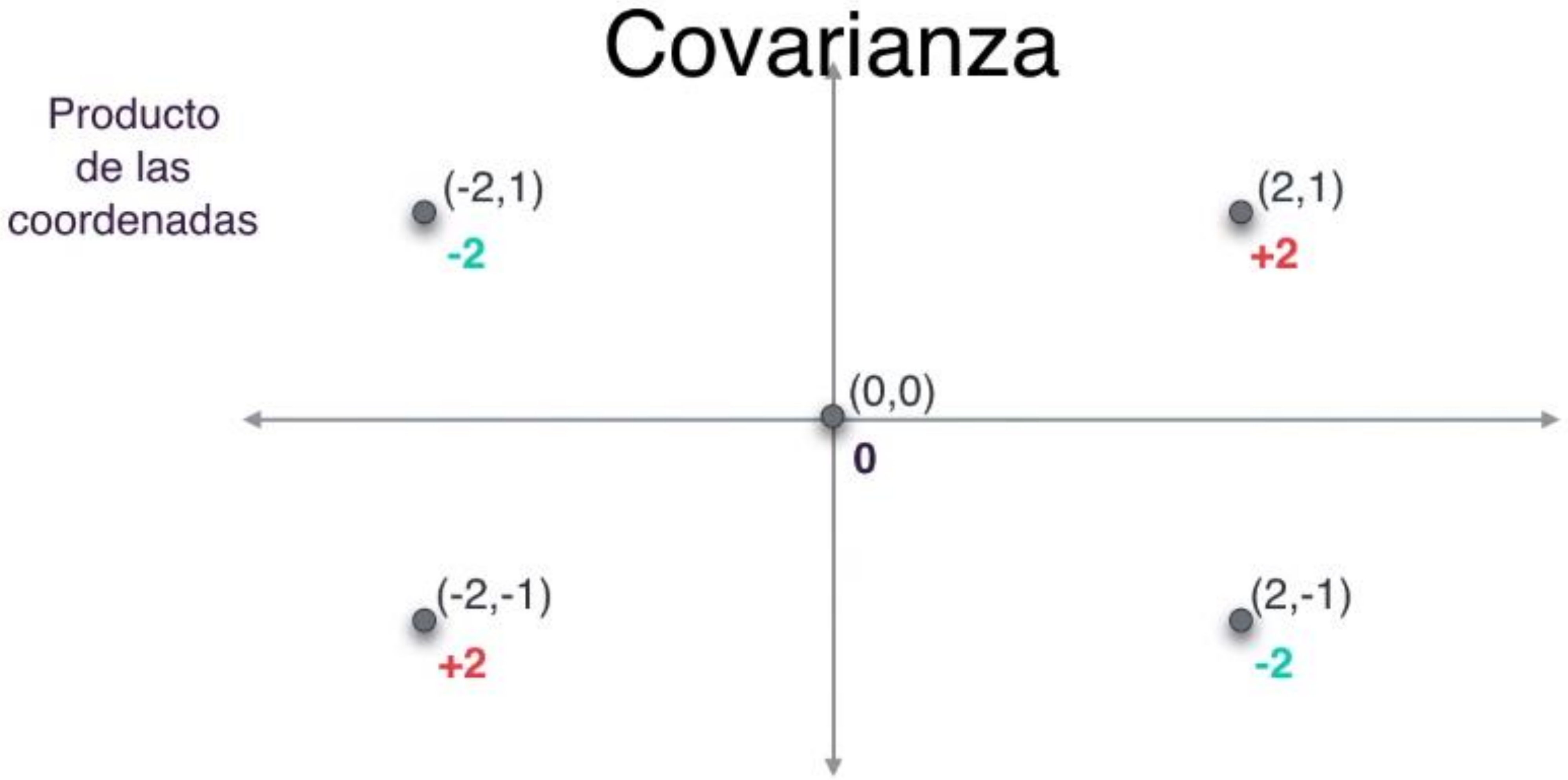


$$\text{x-varianza} = \frac{2^2+0^2+2^2}{3} = 8/3$$

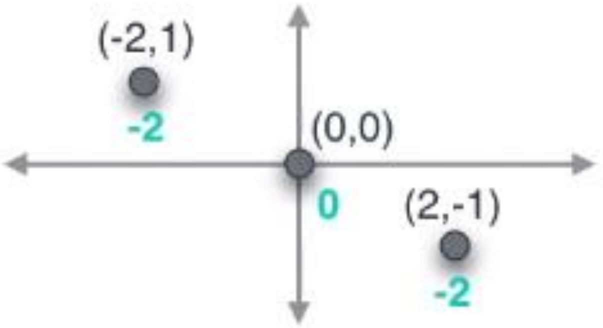
$$\text{y-varianza} = \frac{1^2+0^2+1^2}{3} = 2/3$$

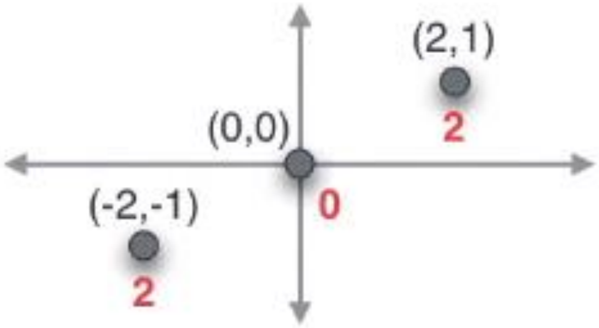
Covarianza




Covarianza
Producto de las coordenadas
(-2,1)
-2
(2,1)
+2
(0,0)
0
(-2,-1)
+2
(2,-1)
-2

Covarianza

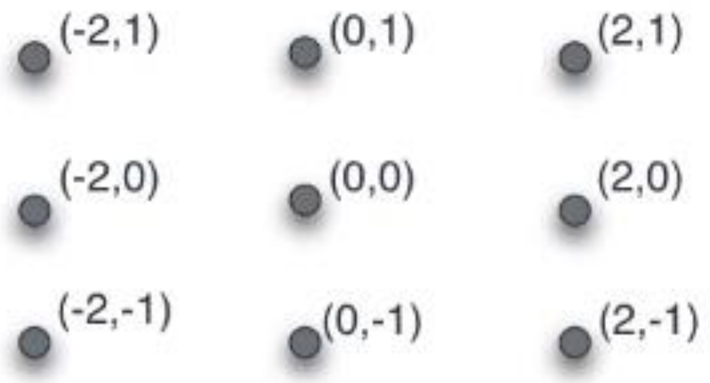




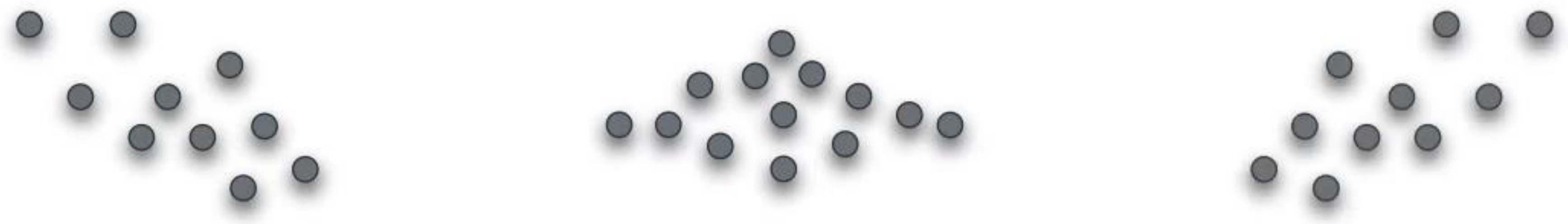
$$\text{covarianza} = \frac{(-2) + 0 + (-2)}{3} = -4/3$$

$$\text{covarianza} = \frac{2 + 0 + 2}{3} = 4/3$$

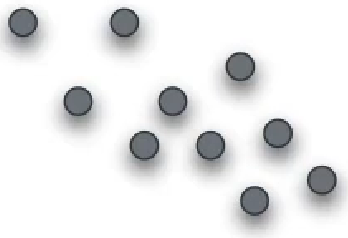
Covarianza



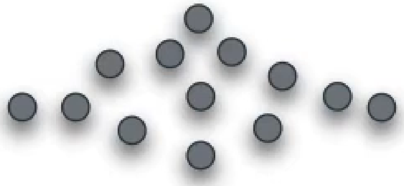
Covarianza



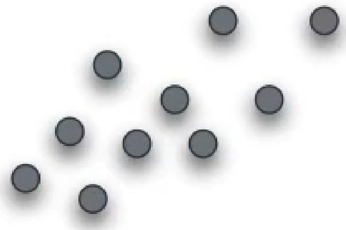
Covarianza



covarianza negativa

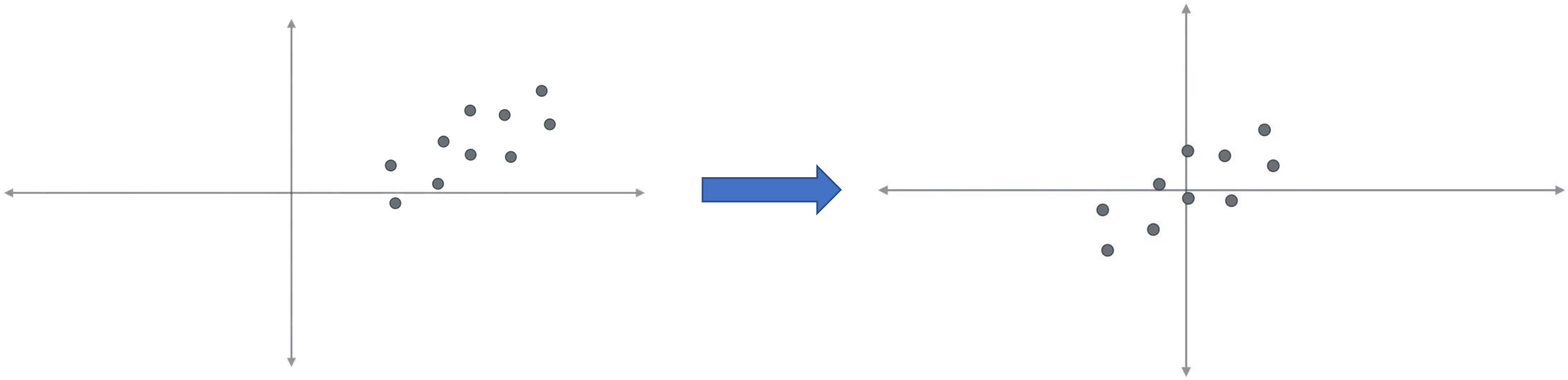


covarianza cero (o muy pequena)

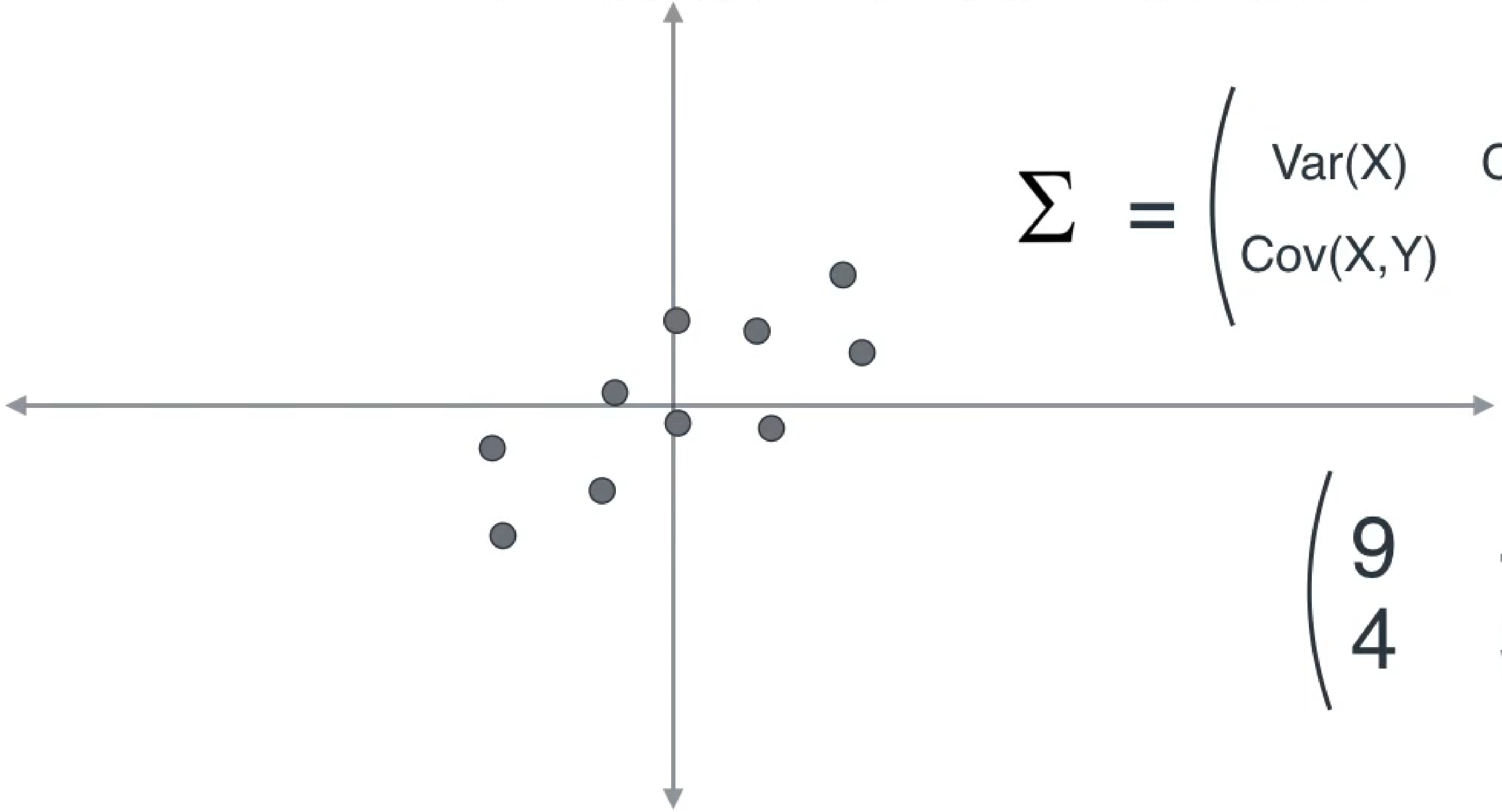


covarianza positiva

Valores y vectores propios (eigenvalues and eigenvectors)



Matriz de covarianza



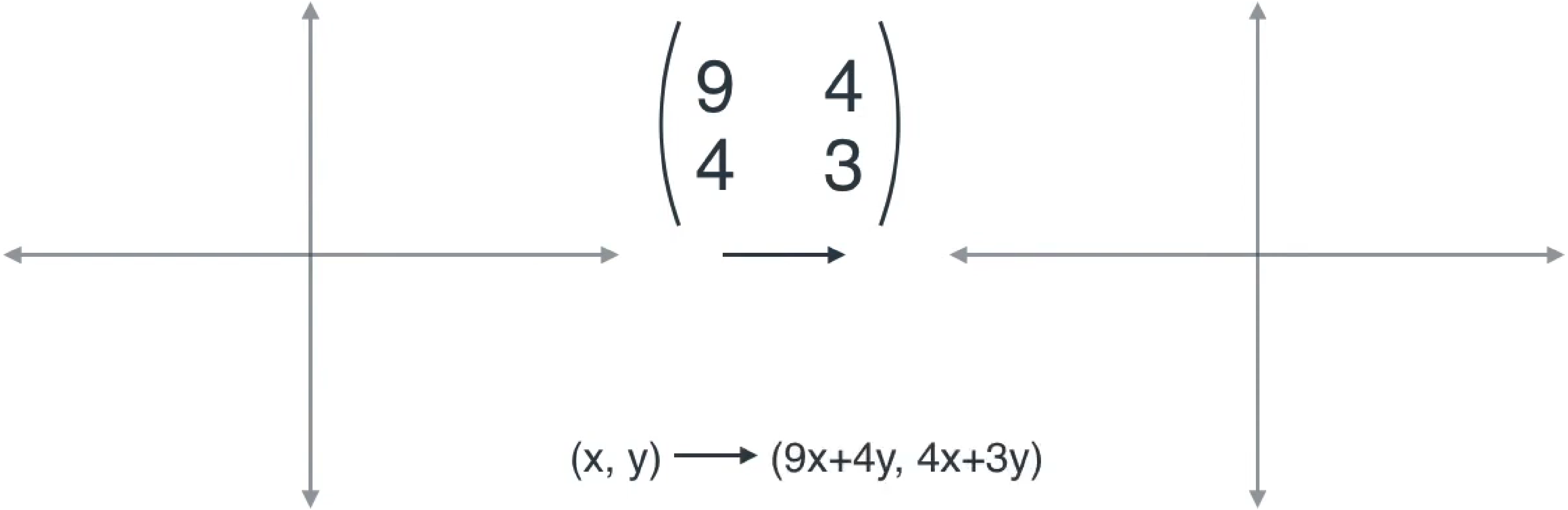
$$\Sigma = \begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix}$$

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

Transformaciones lineales

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

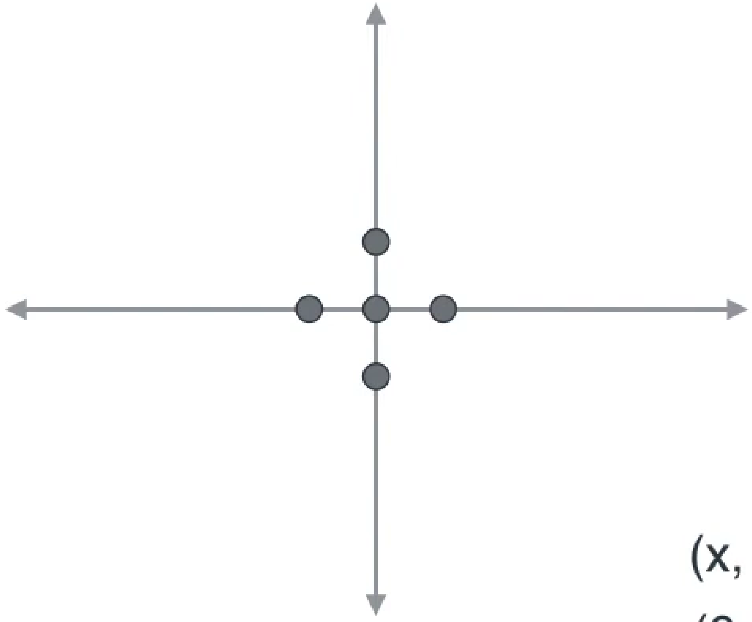
Transformaciones lineales



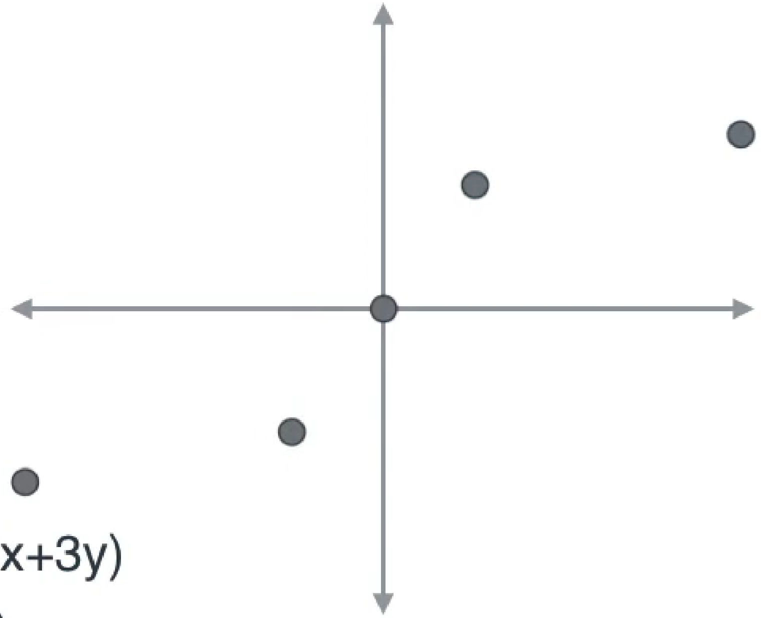
$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

$$(x, y) \longrightarrow (9x+4y,\ 4x+3y)$$

Transformaciones lineales



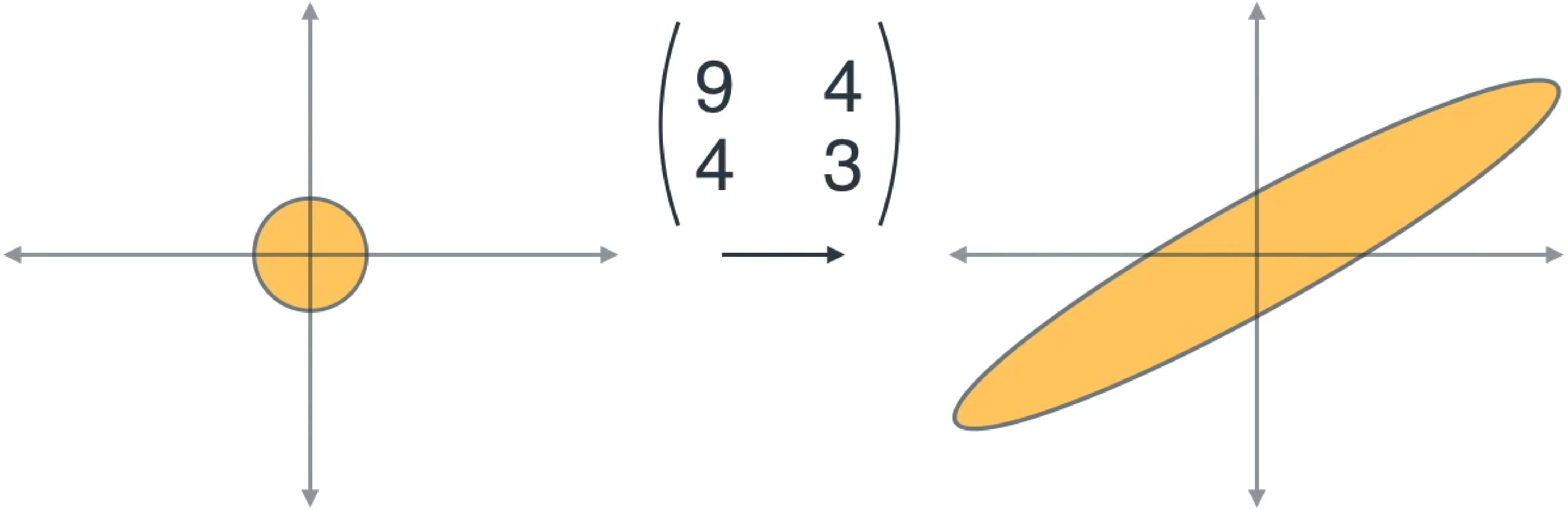
$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix} \longrightarrow$$



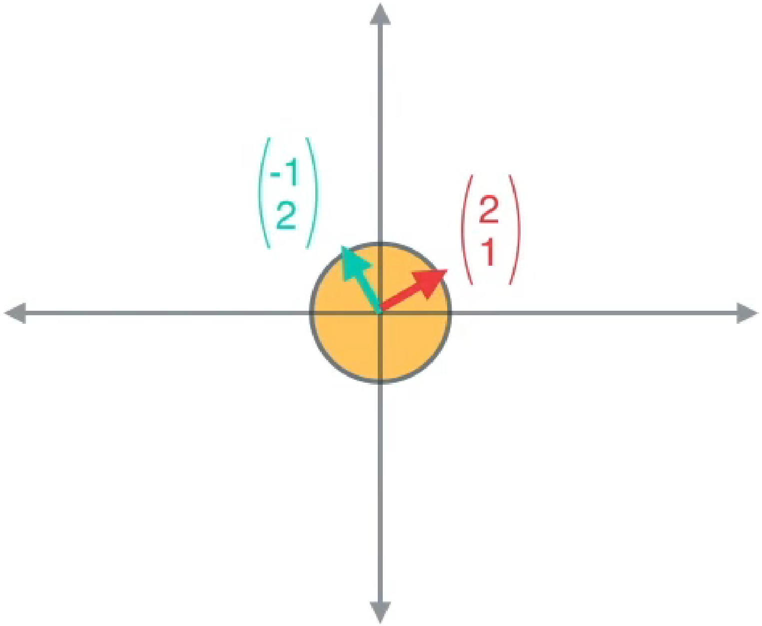
$(x, y) \longrightarrow (9x+4y, 4x+3y)$

(0,0)	(0,0)
(1,0)	(9,4)
(0,1)	(4,3)
(-1,0)	(-9,-4)
(0,-1)	(-4,-3)

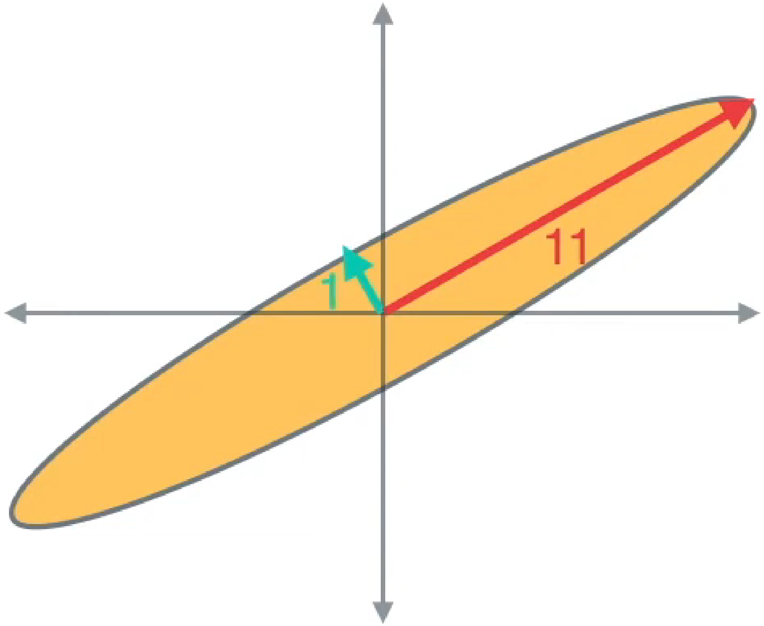
Transformaciones lineales



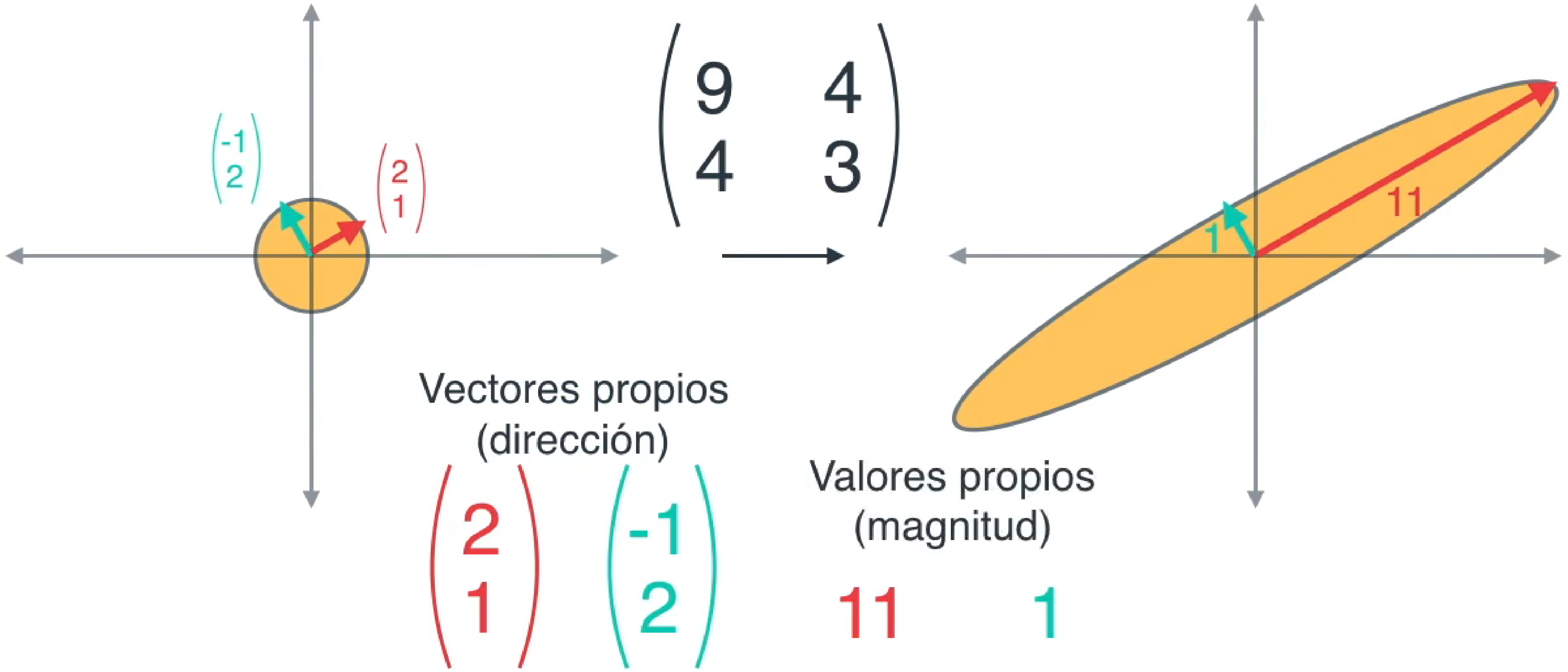
Transformaciones lineales



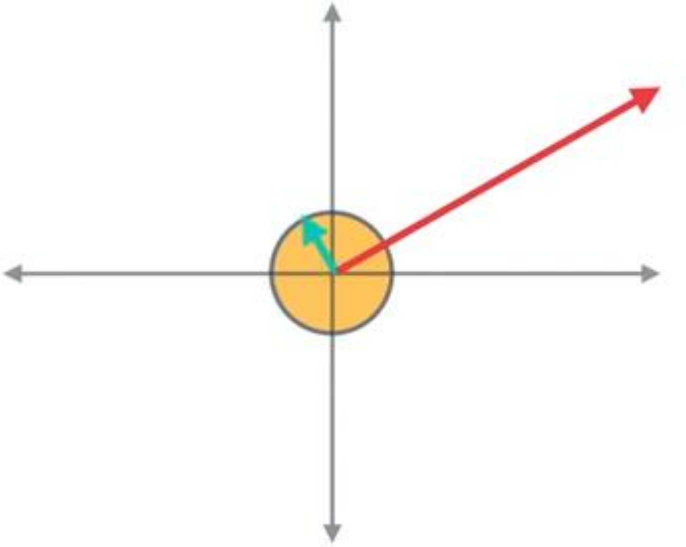
$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix} \longrightarrow$$



Transformaciones lineales



Transformaciones lineales

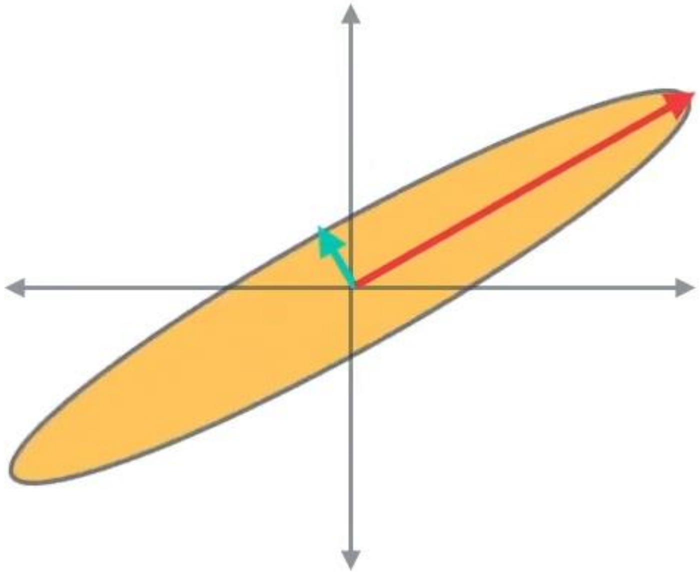


Vectores propios
(dirección)

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

Valores propios
(magnitud)

11 1



Valores y vectores propios

Vector propio

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix} \mathrm{v} = \lambda \, \mathrm{v}$$

Valor propio

Valores propios

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

Polinomio característico

$$\begin{vmatrix} x-9 & -4 \\ -4 & x-3 \end{vmatrix} = (x-9)(x-3) - (-4)(-4) = x^2 - 12x + 11$$

$$= (x-11)(x-1)$$

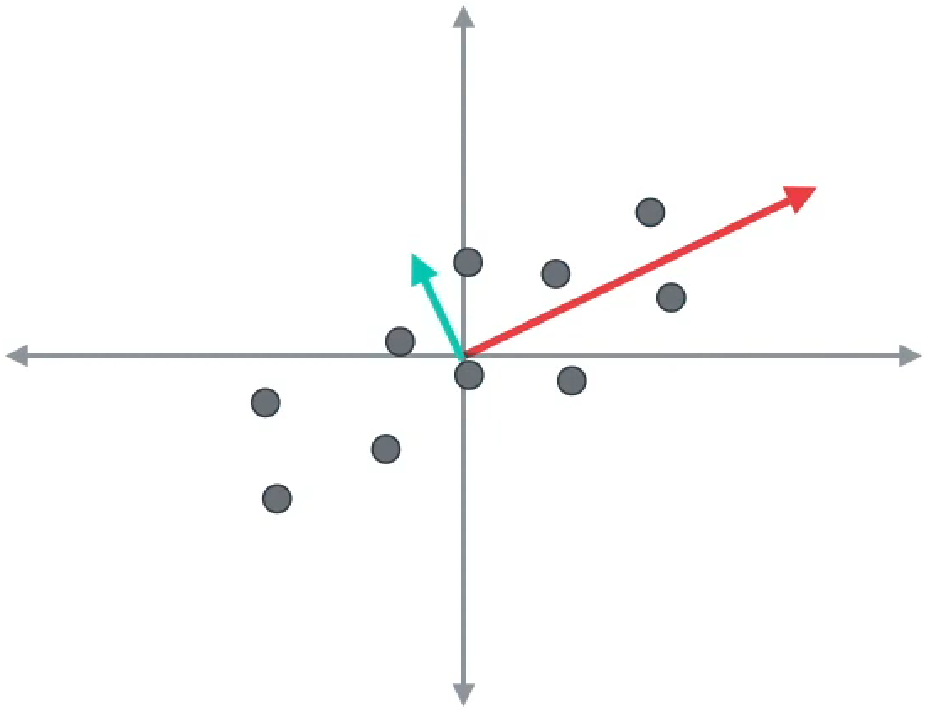
Valores propios 11 y 1

Vectores propios

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = 11\begin{pmatrix} u \\ v \end{pmatrix} \qquad \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = 1\begin{pmatrix} u \\ v \end{pmatrix}$$

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \qquad \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

Análisis de componentes principales (PCA)

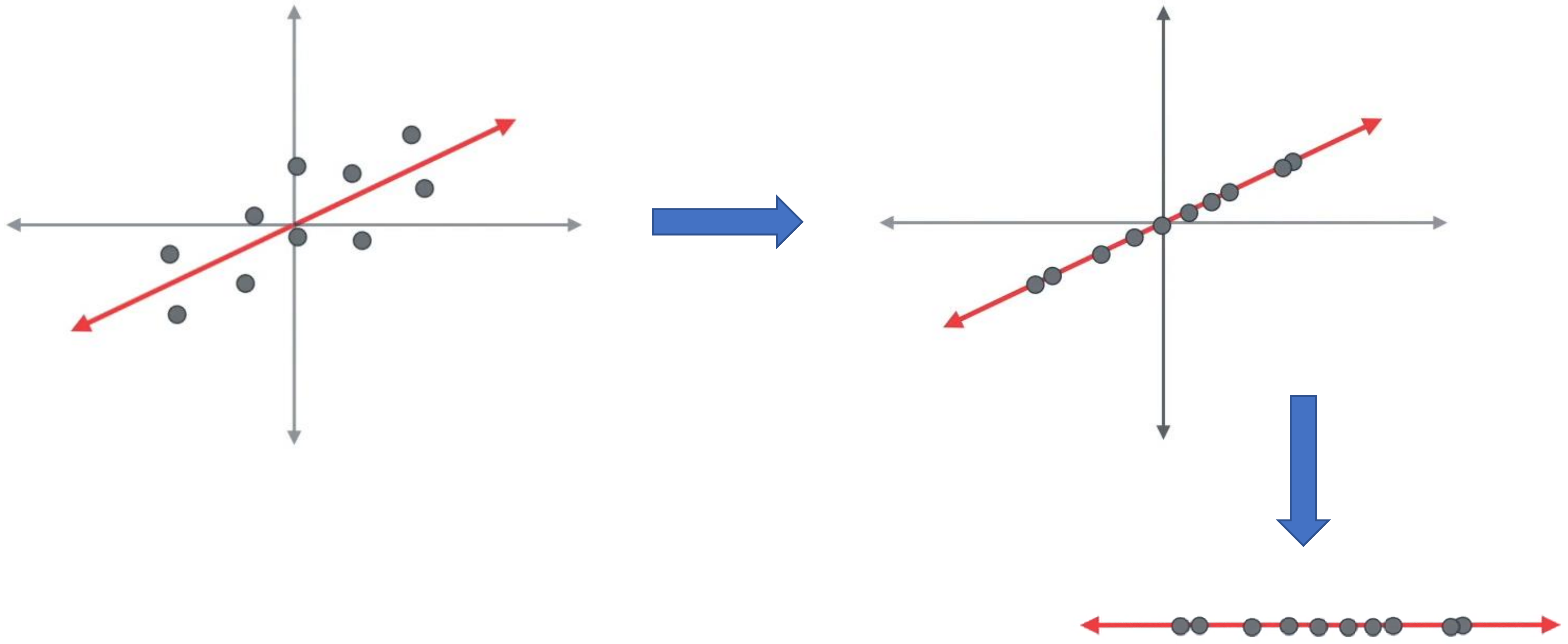


$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

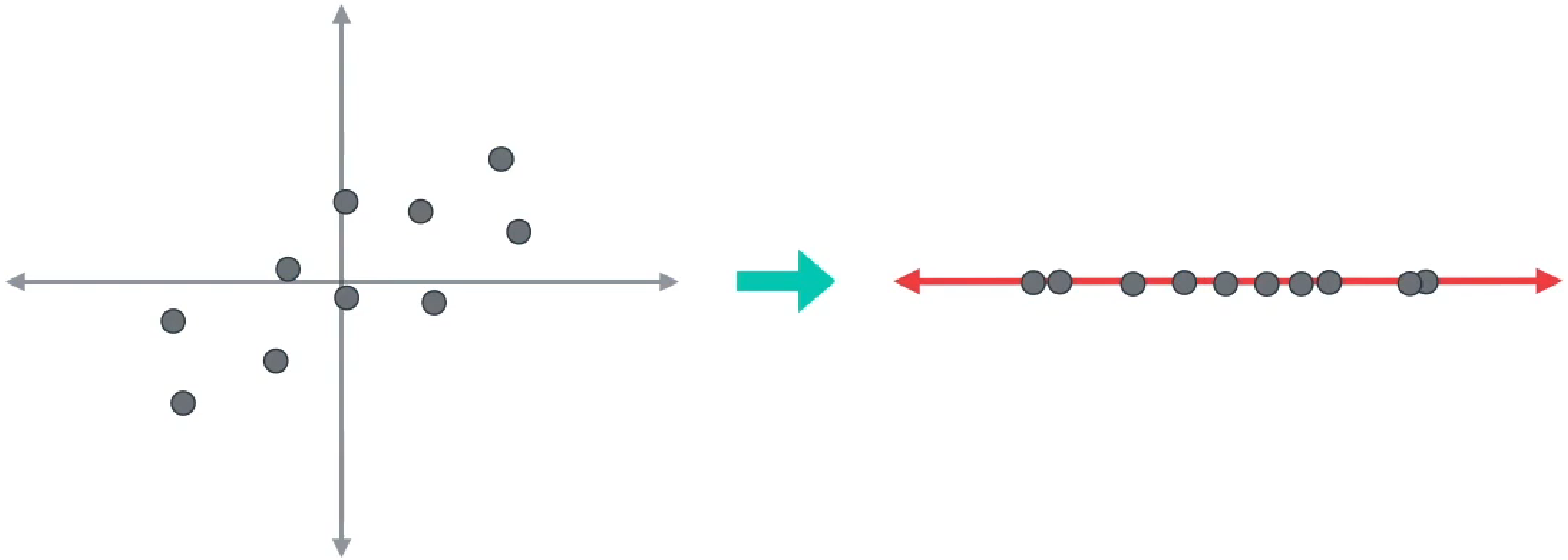
$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$ Vectores propios (dirección)

$11 \quad 1$ Valores propios (magnitud)

Análisis de componentes principales (PCA)



Análisis de componentes principales (PCA)



PCA

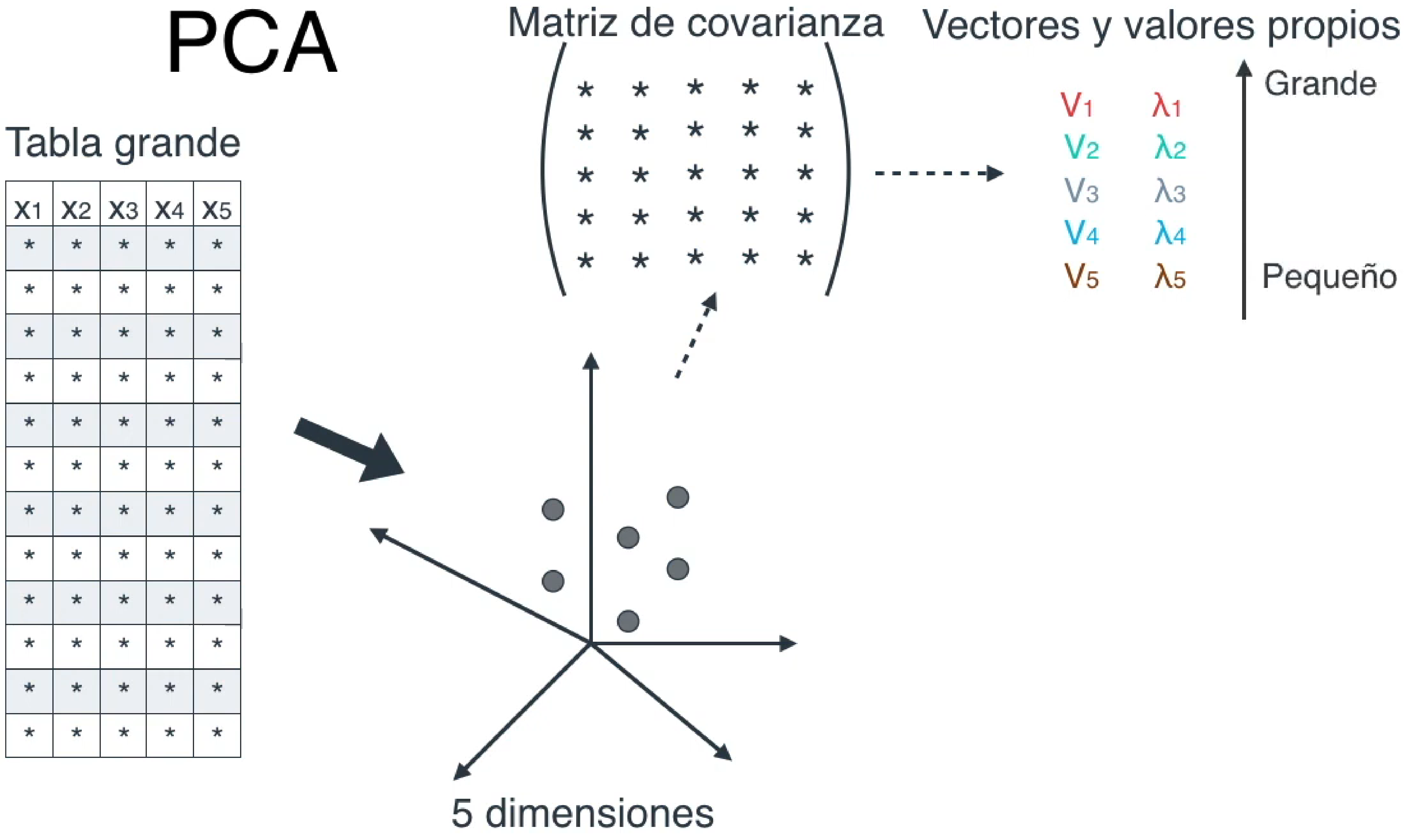
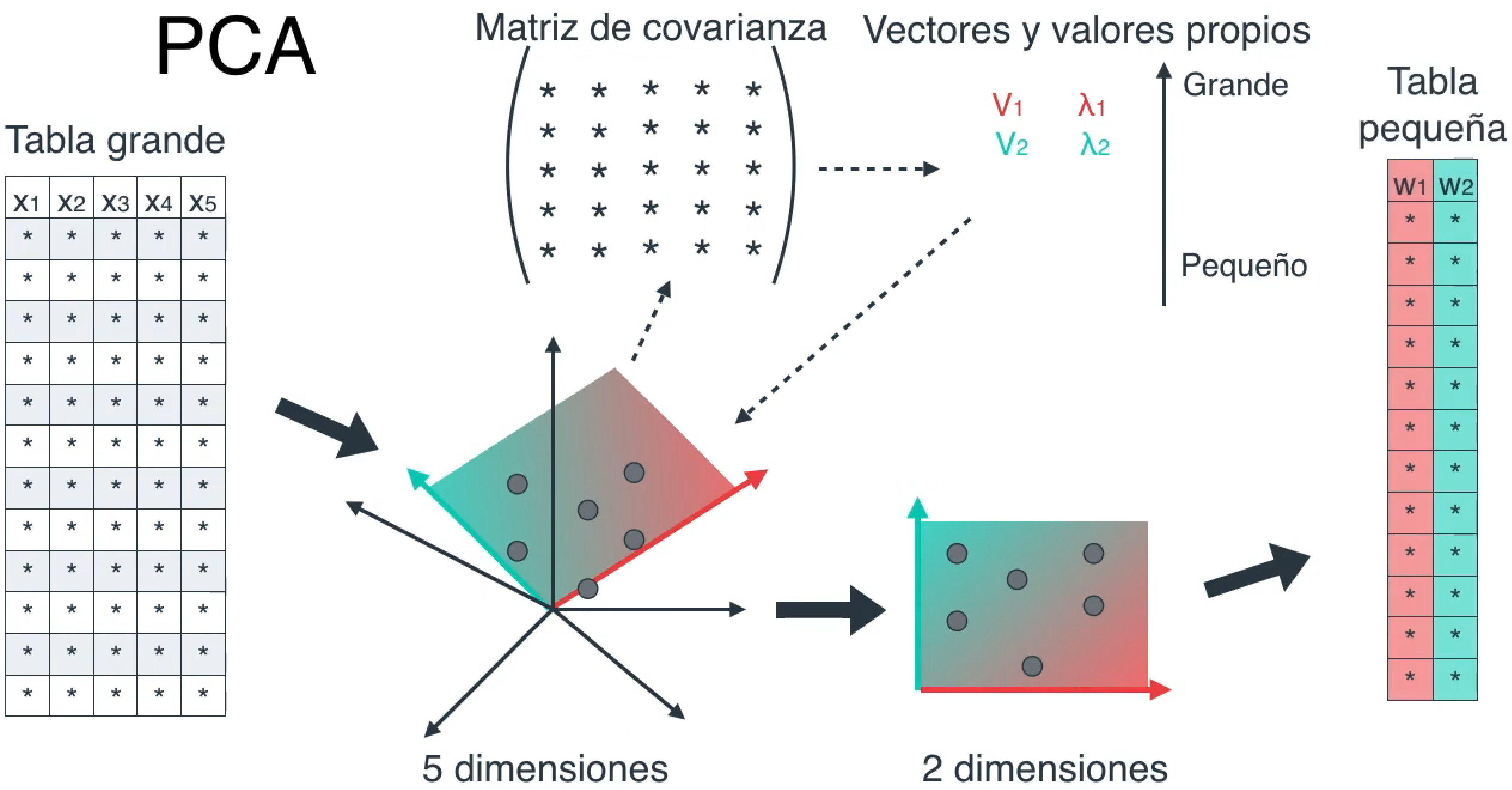


Tabla grande
X1 X2 X3 X4 X5
Matriz de covarianza
Vectores y valores propios
V1 λ1
V2 λ2
V3 λ3
V4 λ4
V5 λ5
Grande
Pequeño
5 dimensiones

PCA



Matriz de covarianza
Vectores y valores propios
Tabla grande
X1 X2 X3 X4 X5
V1 λ1
V2 λ2
Grande
Pequeño
Tabla pequeña
W1 W2
5 dimensiones
2 dimensiones

Transformación de variables

Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6

Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6

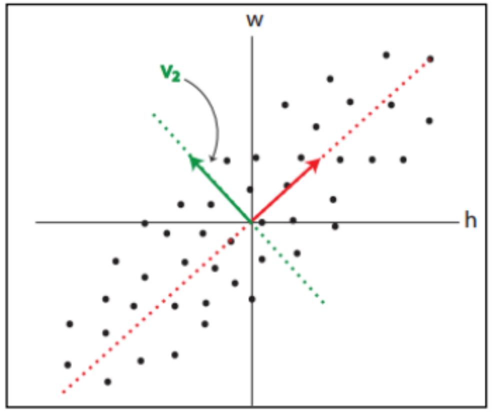
- Los primeros componentes reúnen la mayor cantidad de varianza de los datos originales.
- Los componentes son combinaciones lineales de las variables originales.

$$Comp_1 = w_{11}Var_1 + w_{12}Var_2 + \cdots + w_{1n}Var_n$$

Conceptos

1. Dirección
Vectores propios

2. Magnitud
Valores propios



$$\mathbf{v}_1 = \begin{pmatrix} 0.7 \\ 0.7 \end{pmatrix} = 0.7\mathbf{h} + 0.7\mathbf{w}$$

3. Varianza explicada de los datos
- Absoluta
- Relativa. Peso relativo del valor propio sobre la suma de todos los valores propios

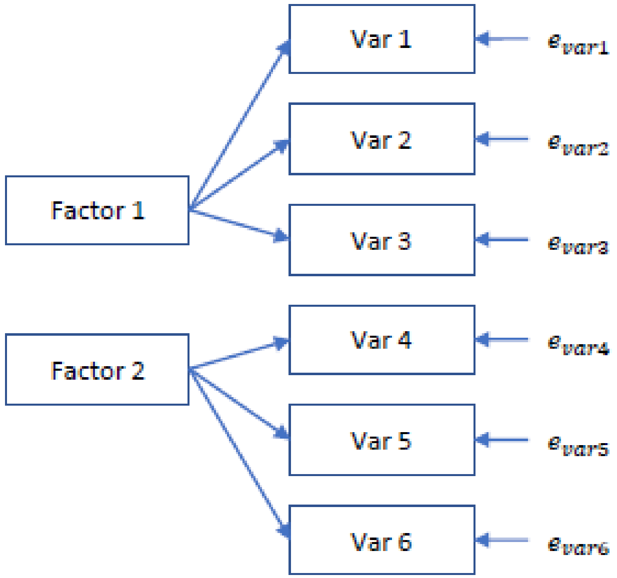
4. Loadings (Carga del factor)
Los coeficientes de la combinación lineal se denomina cargas

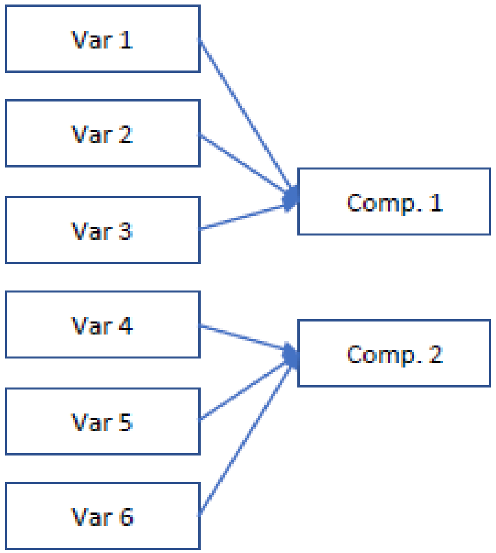
5. Varianza explicada de cada variable
Comunalidades: Suma de cargas al cuadrado

Análisis Factorial

El objetivo es develar la estructura de los datos. El fin último generalmente es detectar redundancia de variables (o casos) y, si se desea, realizar reducción de variables independientes (o casos). Otro posible objetivo es detectar factores latentes.

Análisis Factorial Vs Componentes Principales

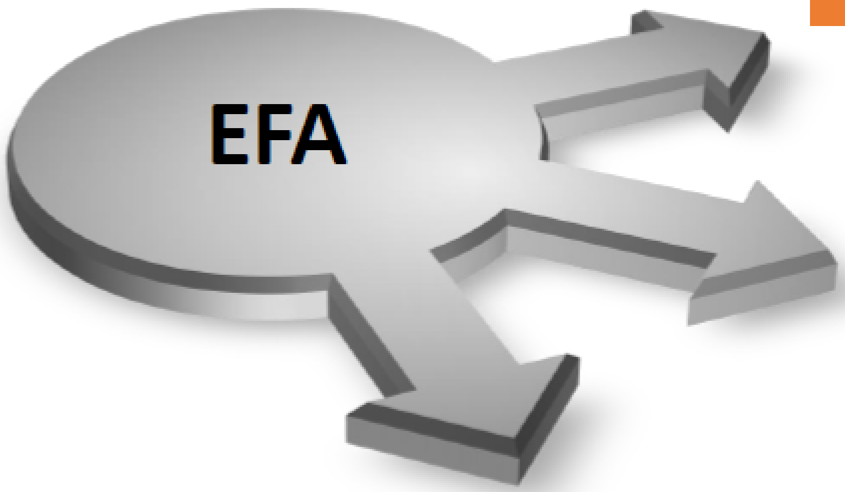




EFA: Variables observadas son combinaciones lineales de los factores latentes.

PCA: Los componentes son combinaciones lineales de las variables observadas.

Análisis Factorial



Objetivo:

- Determinar factores latentes a las variables observadas
- Disminución de dimensiones

Técnica de extracción:

- Componentes principales
- Mínimos cuadrados: entre matriz de correlación y matriz reproducida
- Máxima verosimilitud: supuesto de normalidad multivariada

Rotación:

Se rotan los factores latentes para mejorar su interpretación.

Supuestos

Correlación Lineal

- Las Relaciones que se prueban únicamente son lineales
- Pueden utilizarse otro tipo de coeficientes de correlación

Naturaleza de las variables

- Deben ser variables escalares pero puede adaptarse para variables ordinales

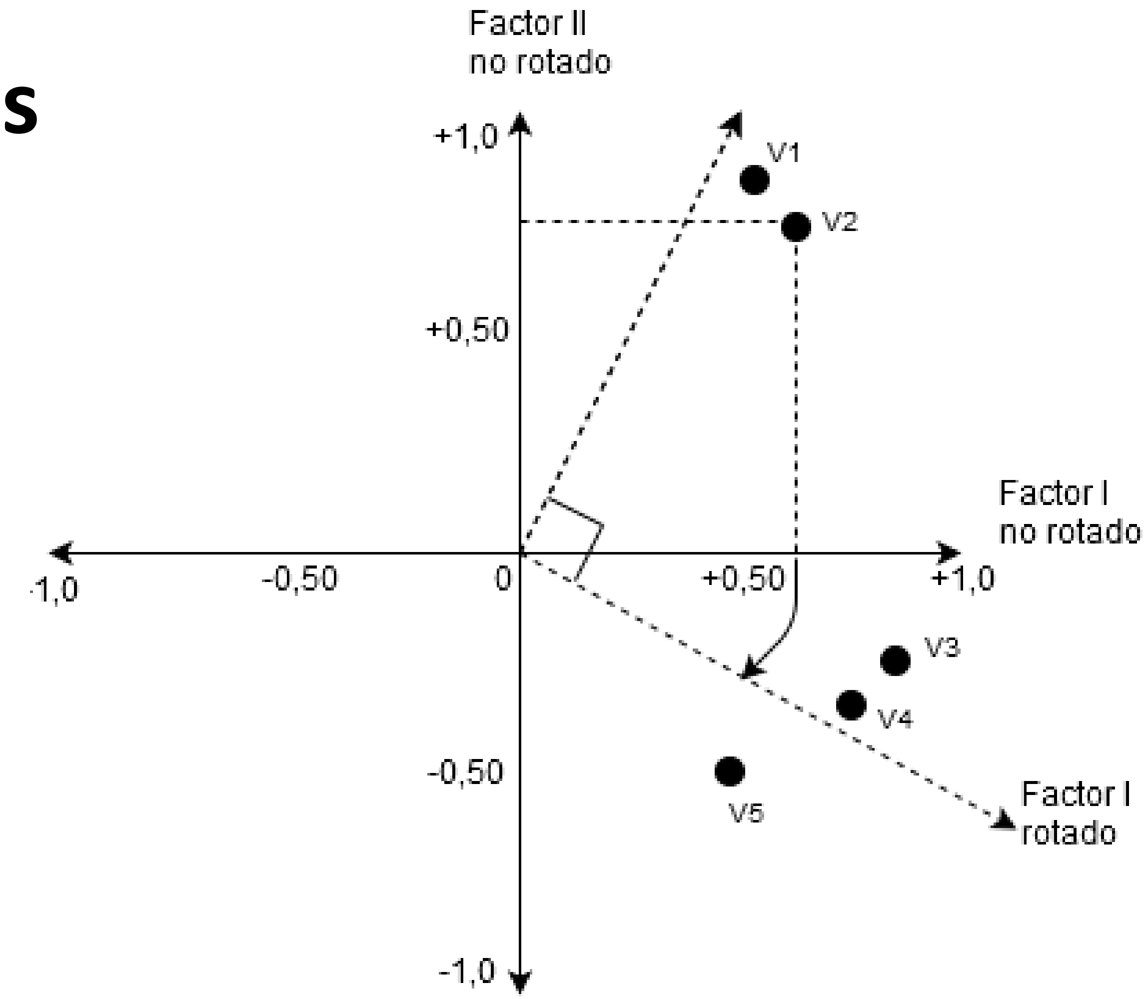
Factorabilidad

- Existe correlación entre las variables
- Colinearidad. KMO, Barlett

Tamaño de muestra

- N/k>20:1

Rotación de ejes



Rotaciones: Varimax - Quartimax

Original

	Factor 1	Factor 2
Variable 1	0.5	0.4
Variable 2	-0.2	0.5
Variable 3	0.3	0.4
Variable 4	0.6	0.2

Varimax

	Factor 1	Factor 2
Variable 1	0.8	0.4
Variable 2	-0.2	0.3
Variable 3	0.3	0.9
Variable 4	0.9	0.2

Quartimax

	Factor 1	Factor 2
Variable 1	0.6	0.2
Variable 2	-0.1	0.6
Variable 3	0.2	0.65
Variable 4	0.7	0.05

Las cargas factoriales no son representativas de valores verdaderos, solo ilustrativas del concepto

Data Understanding/Preparation

- **Estandarizar variables**
  - Vuelve comparables las magnitudes de las variables
- **Análisis de correlaciones**
  - Permite darse una idea de la estructura de correlación en los datos
  - Da cierta forma valida la pertinencia de la reducción de dimensiones
  - Puede hacerse sobre:
    - ✓ **Mapa de calor de matriz de correlaciones/covarianzas:** debe percibirse cierta correlación entre las variables
    - ✓ **Prueba de esfericidad de Barlet:** contrasta la hipótesis de que la matriz de covarianzas es la matriz identidad
    - ✓ **KMO:** Compara correlaciones vs las correlaciones parciales. Regla del pulgar KMO>0.5

Aplicaciones de PCA/EFA

- **Reducción de dimensiones**
  - Volver escalables algunos algoritmos aplicados dentro de Analítica
- **Identificar factores latentes**
  - Aplicados usualmente dentro del campo de la psicología
  - A través de información "indirecta" ayudan, de cierta manera, a la medición de variables que no se pueden observar directamente: felicidad, inteligencia, satisfacción, etc.
  - Permite mejorar la comprensión de grandes volúmenes de variables
  - Puede apoyar técnicas de clustering