



IIC 2433 Minería de Datos

<https://github.com/marcelomendoza/IIC2433>

- OUTLINE -

¿Qué vamos a ver?

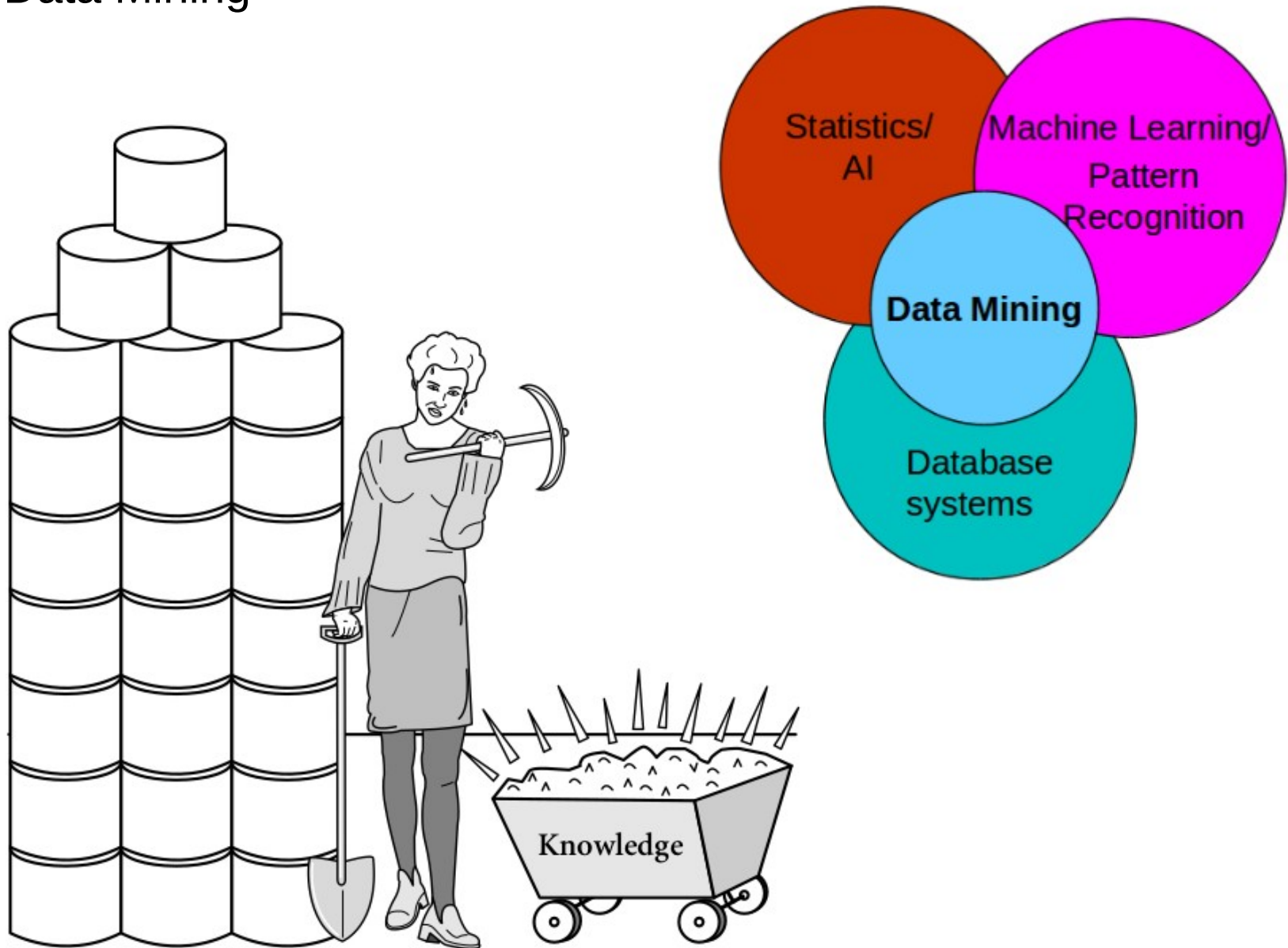
PCA, t-SNE, UMAP, outliers

k-means, HAC, DBSCAN, GMM

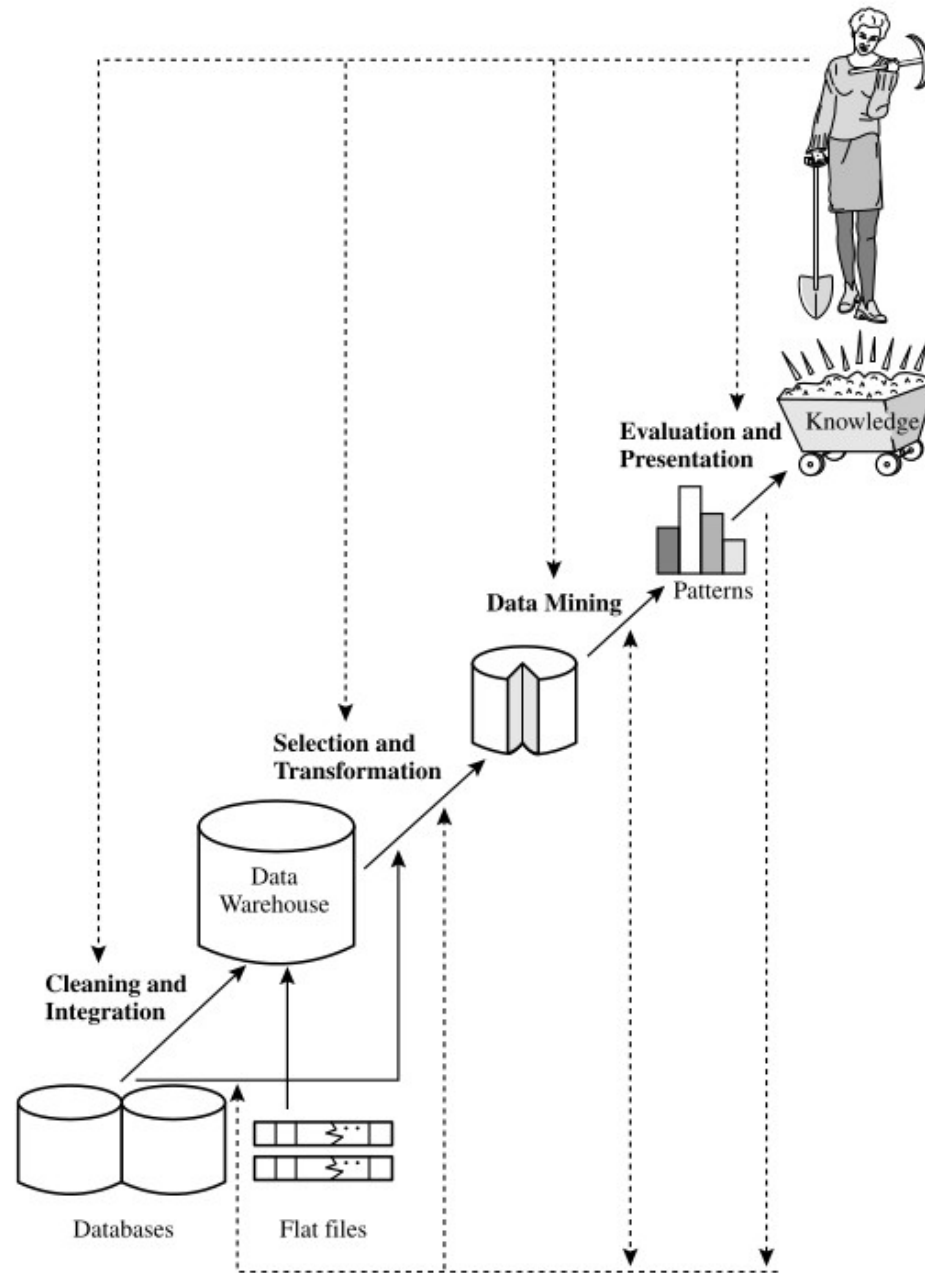
Perceptron, SVM, ensembles

AE. VAE

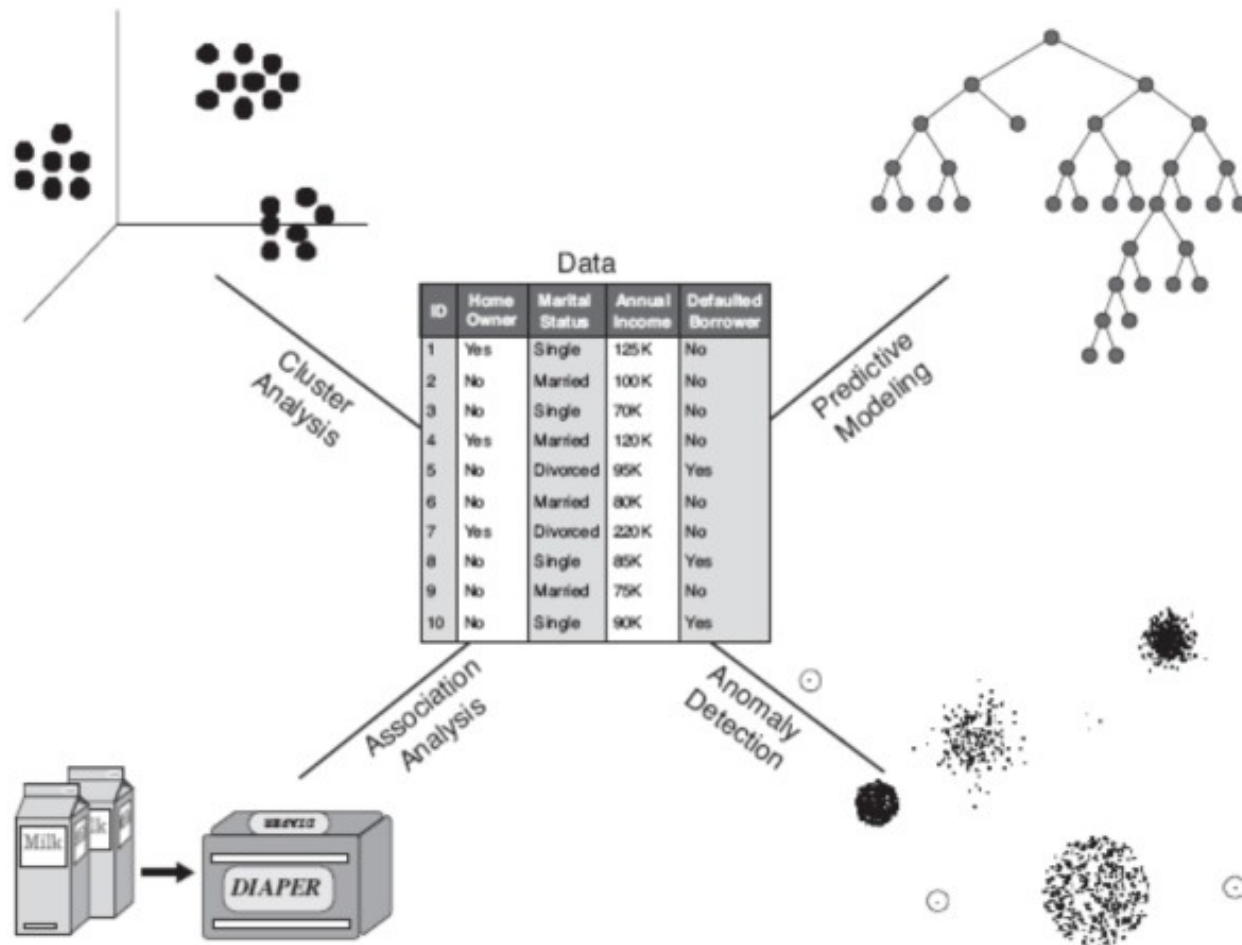
Data Mining



Data Mining



Data Mining



- Preprocesamiento de datos -

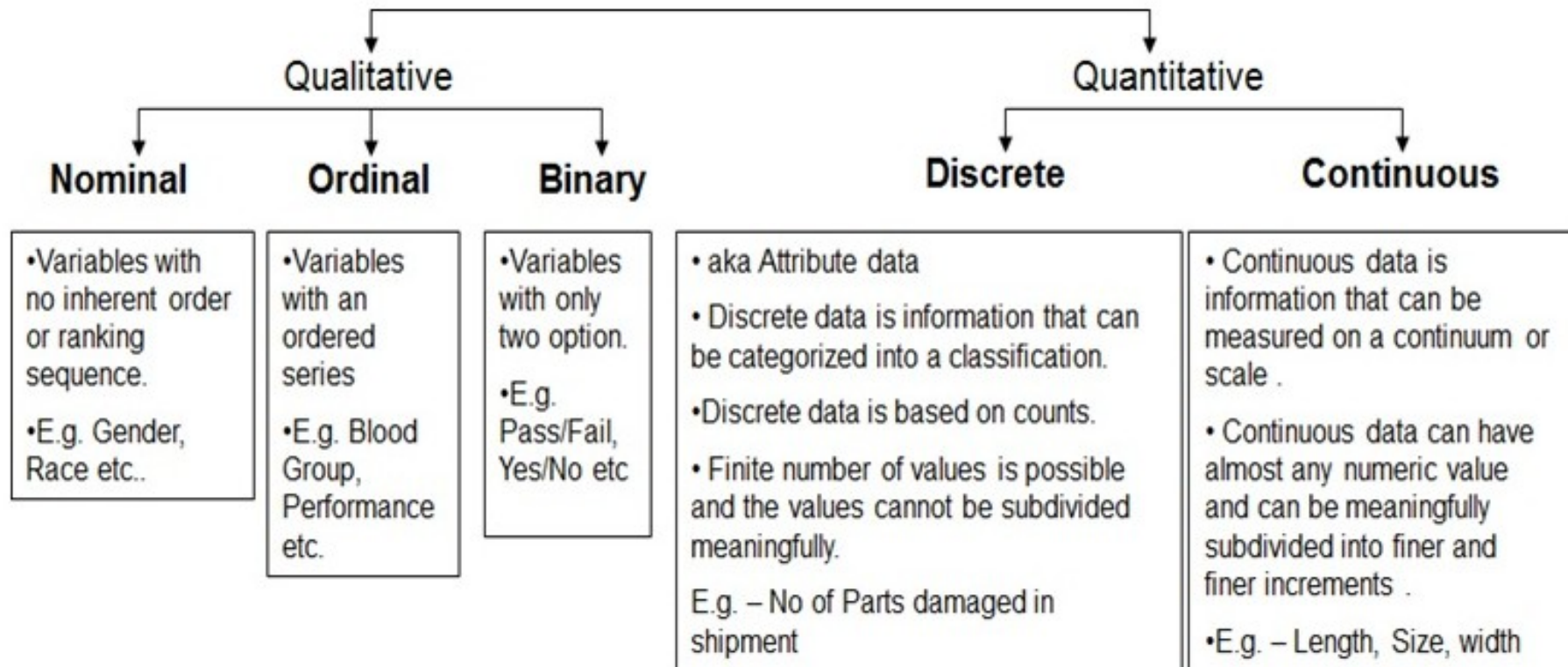
Datos estructurados (tabulares)

	A	B	C	D	E	F	G	H
1	Provincia ▼	Vendedor ▼	Fecha ▼	Nº Pedido ▼	Producto ▼	Unidades ▼	Precio ▼	Importe ▼
2	León	Marío	28/09/2021	105090	Mezcla Gumbo del chef Anton	13	21,35	277,55
3	Burgos	Almudena	04/05/2021	104987	Chocolate Schoggi	17	43,9	746,3
4	Almería	Susana	23/02/2021	104947	Crema de queso Fløtemys	19	21,5	408,5
5	Barcelona	Susana	12/10/2021	105570	Langostinos tigre Carnarvon	4	62,5	250
6	Lleida	Julio	18/05/2021	105398	Algas Konbu	11	6	66
7	Navarra	Almudena	27/07/2021	105057	Queso de cabra	19	2,5	47,5
8	Alicante/Alacant	Julio	05/09/2021	104848	Peras secas orgánicas del tío Bob	13	30	390
9	Palencia	Susana	13/03/2021	105361	Empanada de carne	6	32,8	196,8
10	Salamanca	Marío	02/01/2021	105074	Café de Malasia	18	46	828
11	Valladolid	Juan Carlos	24/12/2021	106221	Queso gorgonzola Telino	8	12,5	100
12	Teruel	Juan Carlos	06/07/2021	103595	Sirope de regaliz	18	10	180
13	Cantabria	Juan Carlos	01/02/2021	105704	Postre de merengue Pavlova	2	17,45	34,9
14	Salamanca	Juan Carlos	09/01/2021	104498	Arenque ahumado	3	9,5	28,5
15	Cantabria	Marío	18/03/2021	103536	Cordero Alice Springs	17	39	663
16	Almería	Juan Carlos	16/10/2021	103495	Tarta de azúcar	20	49,3	986
17	Madrid	Susana	19/12/2021	105626	Camembert Pierrot	20	34	680
18	Zaragoza	Juan Carlos	21/09/2021	104600	Queso gorgonzola Telino	7	12,5	87,5
19	Salamanca	Julio	07/05/2021	105770	Pan fino	16	9	144
20	Bizkaia	Amaya	22/06/2021	105112	Salsa de arándanos Northwoods	10	40	400
21	Salamanca	Almudena	12/02/2021	104597	Salsa de pimiento picante de Luisiana	2	21,05	42,1
22	Alicante/Alacant	Amaya	21/07/2021	103017	Licor verde Chartreuse	14	18	252
23	Valladolid	Susana	03/10/2021	103611	Langostinos tigre Carnarvon	10	62,5	625
24	Almería	Almudena	30/12/2021	102654	Carne de cangrejo de Boston	10	18,4	184
25	Castellón/Castelló	Juan Carlos	04/07/2021	104312	Pan fino	12	9	108
26	Cáceres	Marío	23/05/2021	102494	Camembert Pierrot	15	34	510
27	Toledo	Susana	31/03/2021	102717	Postre de merengue Pavlova	3	17,45	52,35
28	Burgos	Almudena	04/03/2021	104718	Chocolate holandés	20	12,75	255

Datos no estructurados



Tipos de características



Características cuantitativas

Scaling:

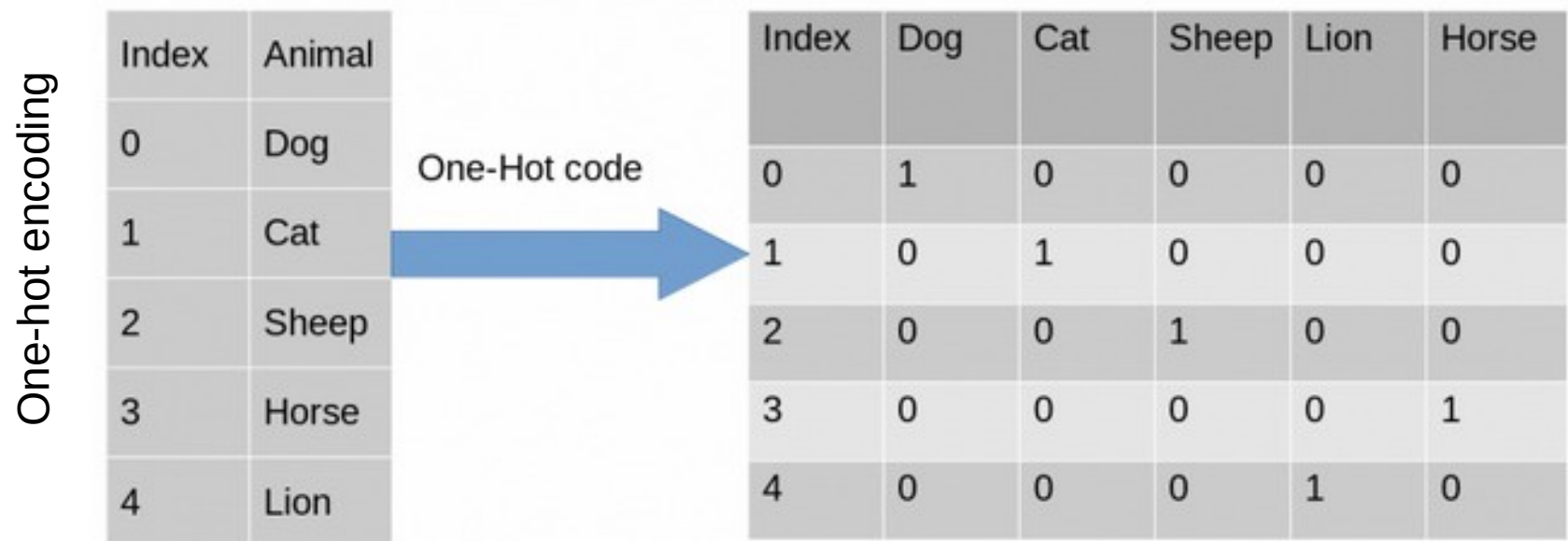
$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \longrightarrow [0, 1]$$

Estandarización:

$$X_{m-norm} = \frac{X - \mu}{X_{max} - X_{min}} \longrightarrow \text{Intervalo centrado en torno de la media}$$

Características cualitativas

Codificación:



La descripción



LA APARIENCIA FÍSICA

SER

alto/a bajo/a delgado/a gordo/a viejo/a joven

rubio/a pelirrojo/a castaño/a

el pelo rubio el pelo pelirrojo el pelo castaño / café el pelo corto el pelo largo

moreno/a

el pelo moreno / negro el pelo gris el pelo blanco el pelo rizado el pelo liso

TENER

los ojos azules los ojos verdes los ojos cafés los ojos negros

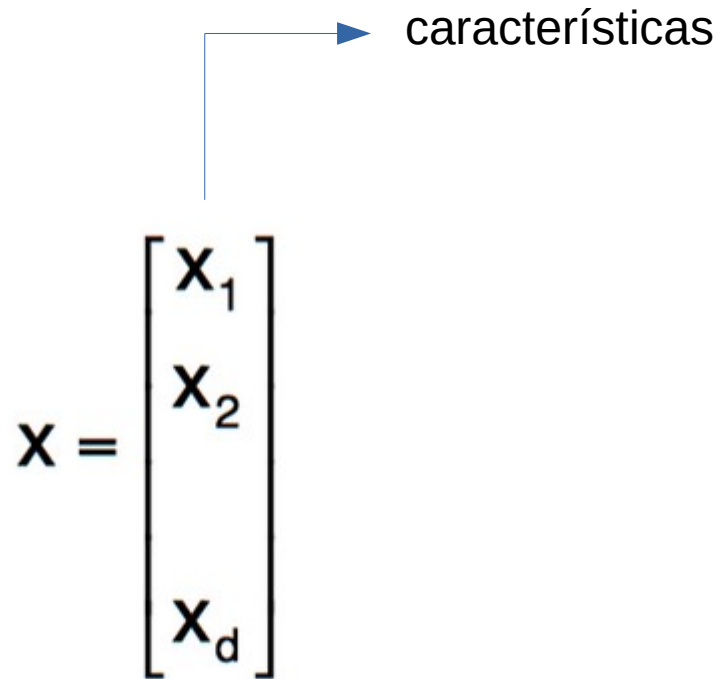
LLEVAR

gafas bigote barba gorra trenzas



Características

Vectores y características

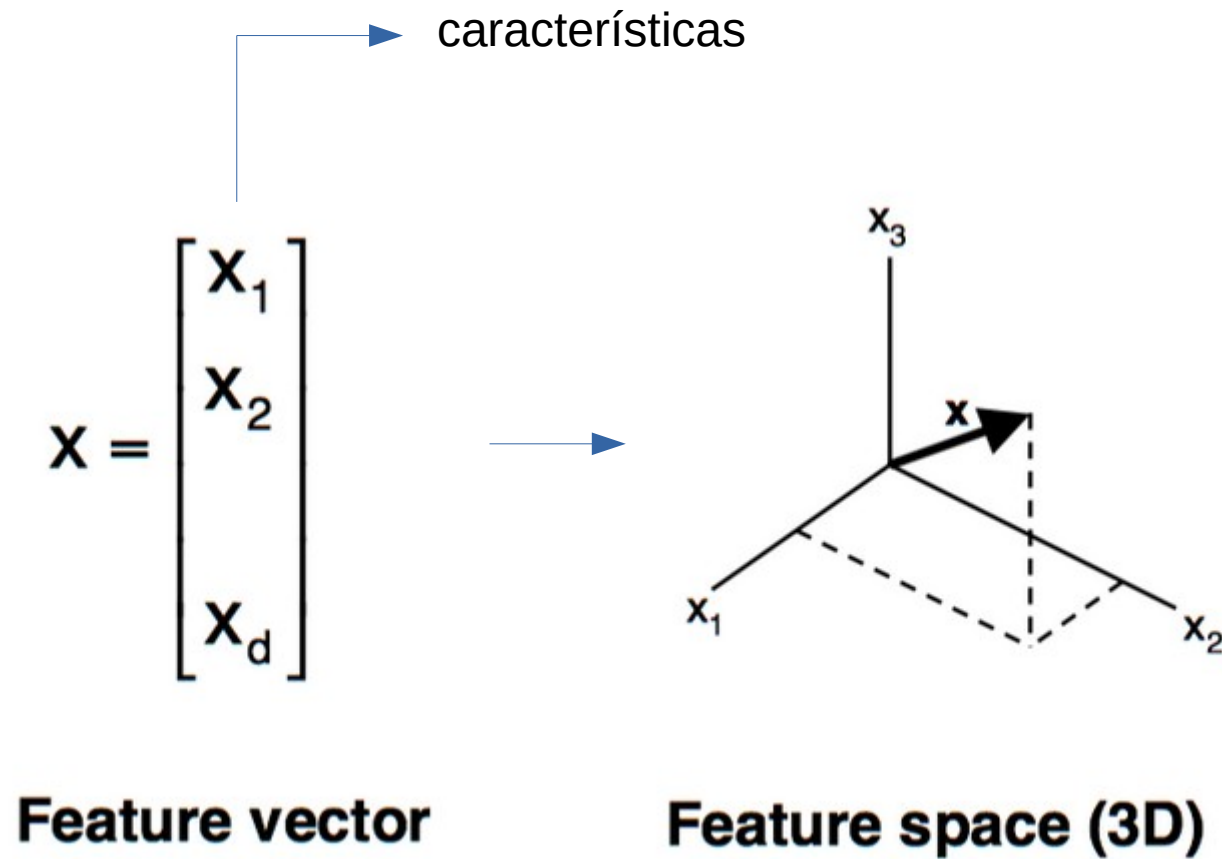


A diagram illustrating a feature vector. A vertical column vector is shown, labeled $\mathbf{x} =$ to its left. The vector contains three elements: x_1 , x_2 , and x_d . A blue arrow points from the first element, x_1 , to the word "características" written to the right of the vector.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

Feature vector

Vectores y características

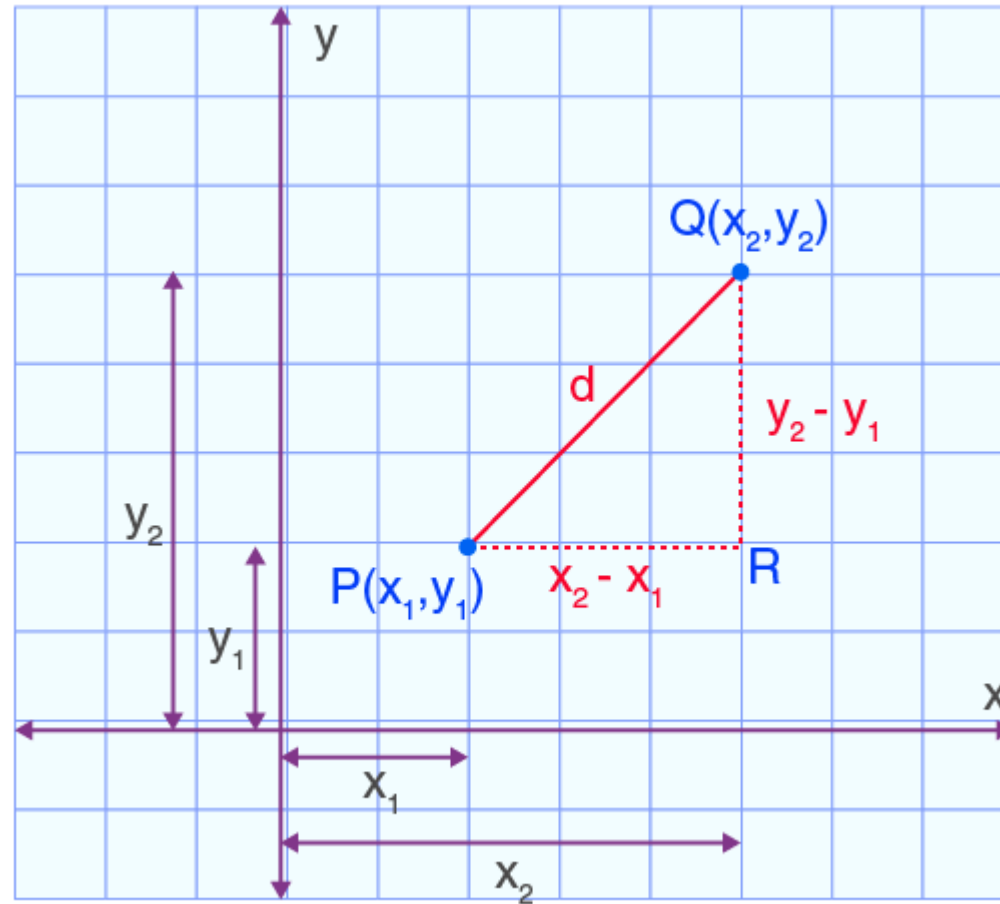


- Distancia y proximidad -



Distancia Euclideana

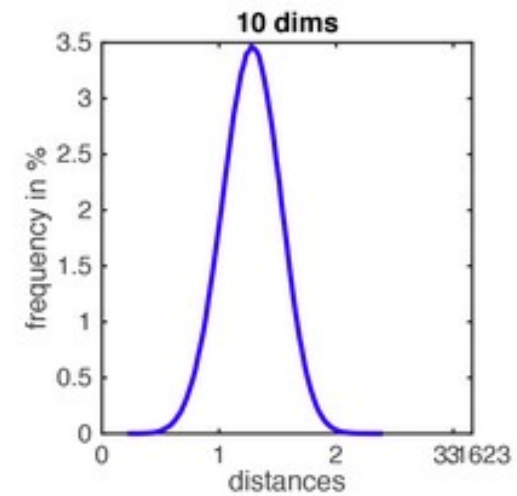
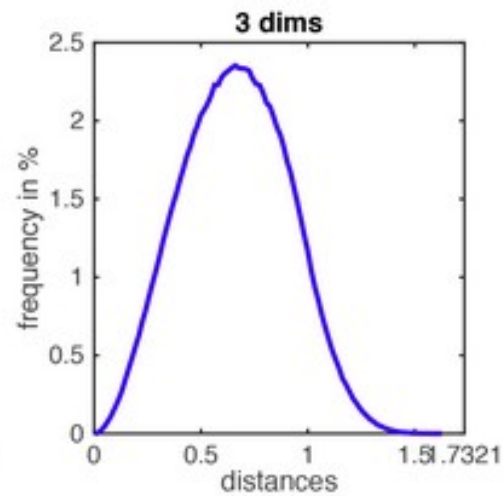
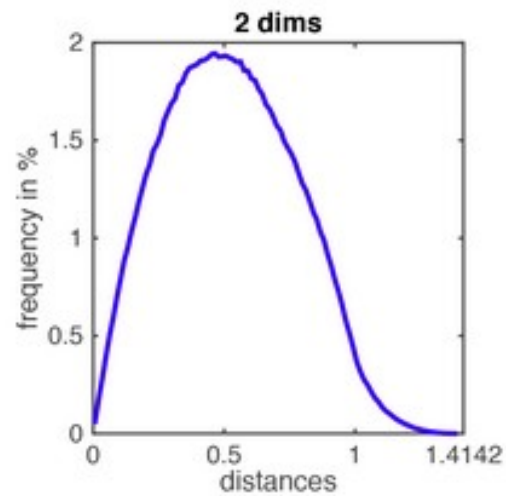
2D:



n-dimensional:
$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$



Distancia Euclidea

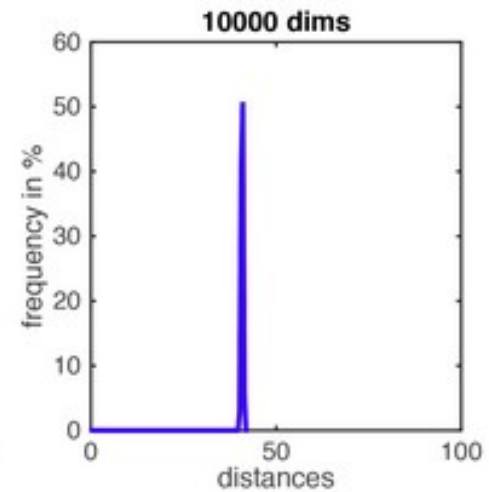
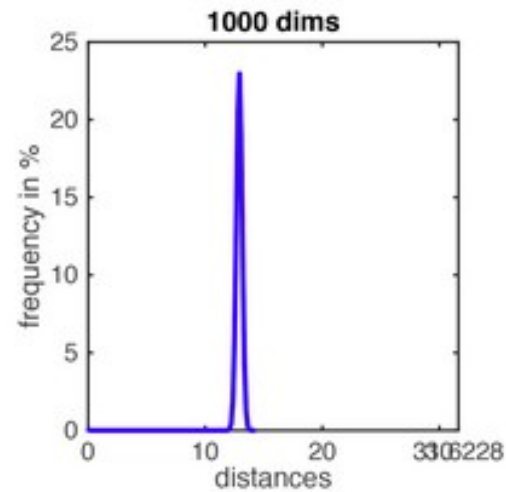
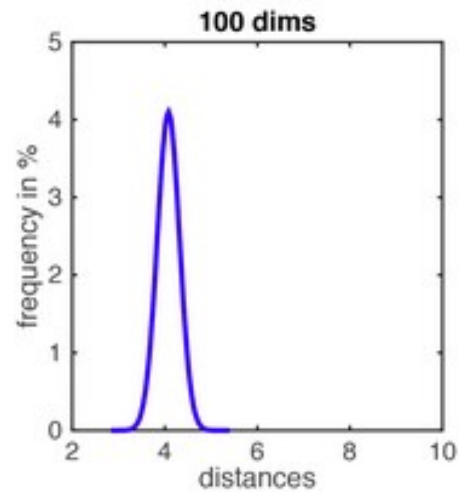


$$\sqrt{2} = 1.41$$

$$\sqrt{3} = 1.73$$

$$\sqrt{10} = 3.16$$

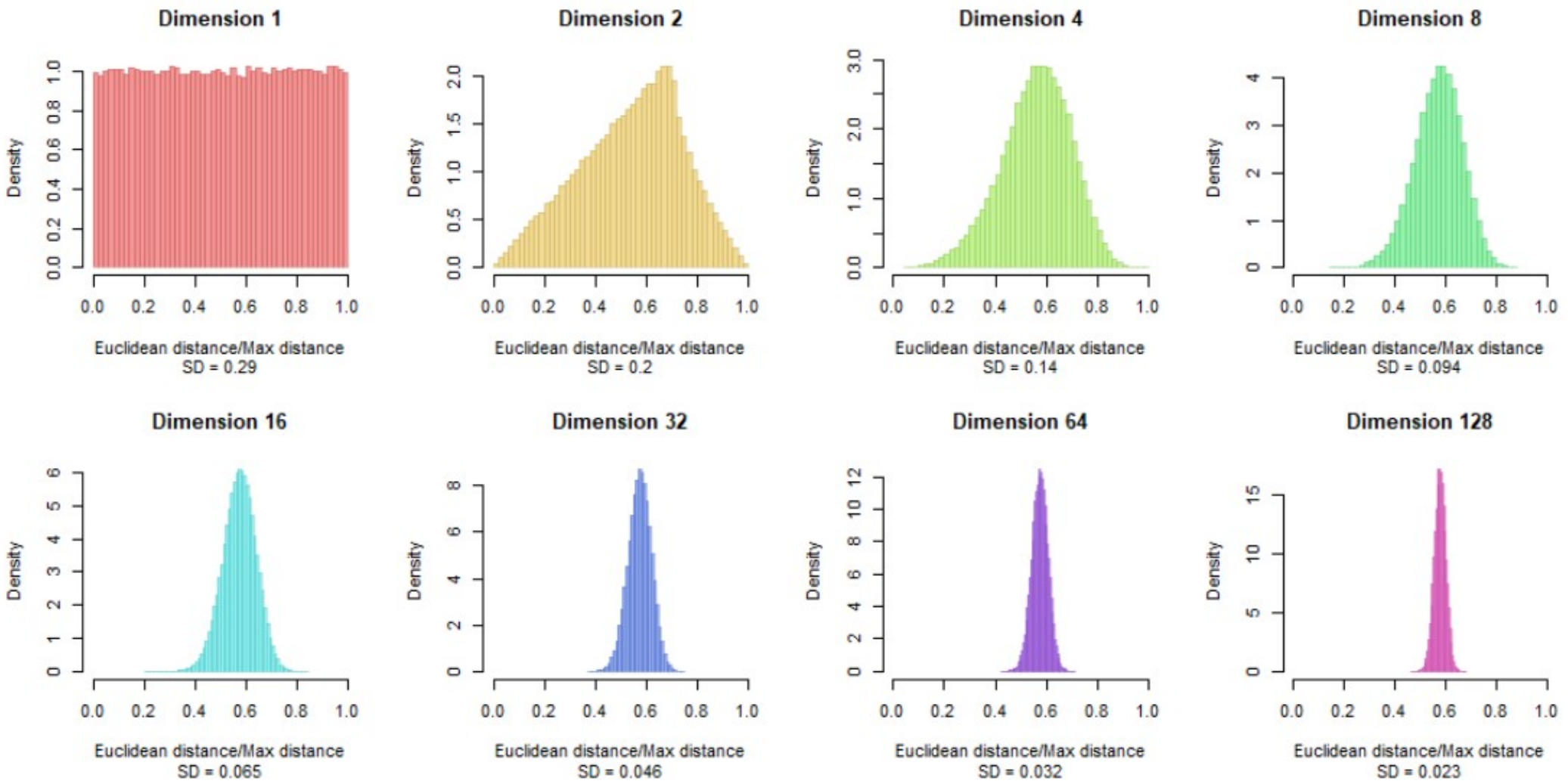
$$\sqrt{1000} = 31.6$$



Maldición de la dimensionalidad

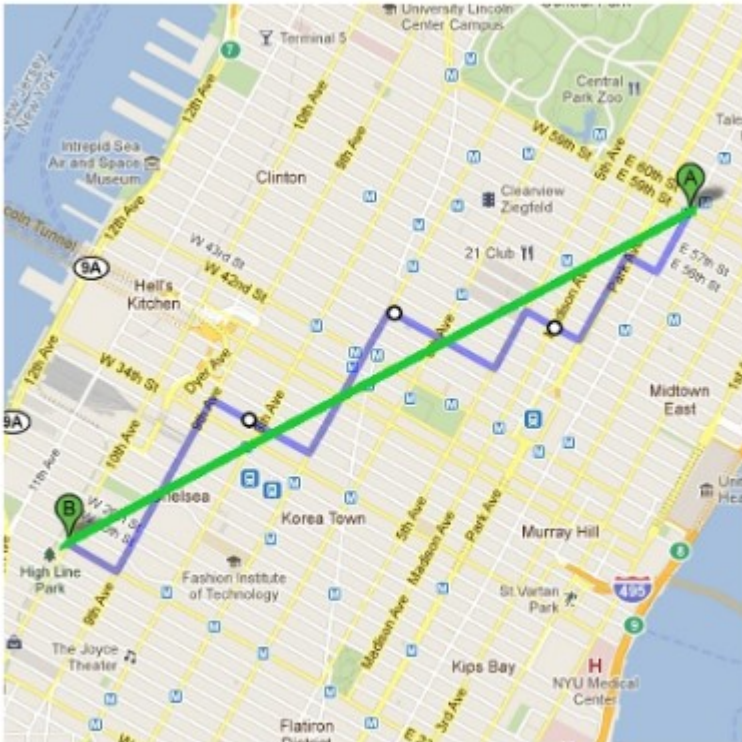


Distancia Euclideana



Maldición de la dimensionalidad (normalizado)

Distancias



Distancia Manhattan

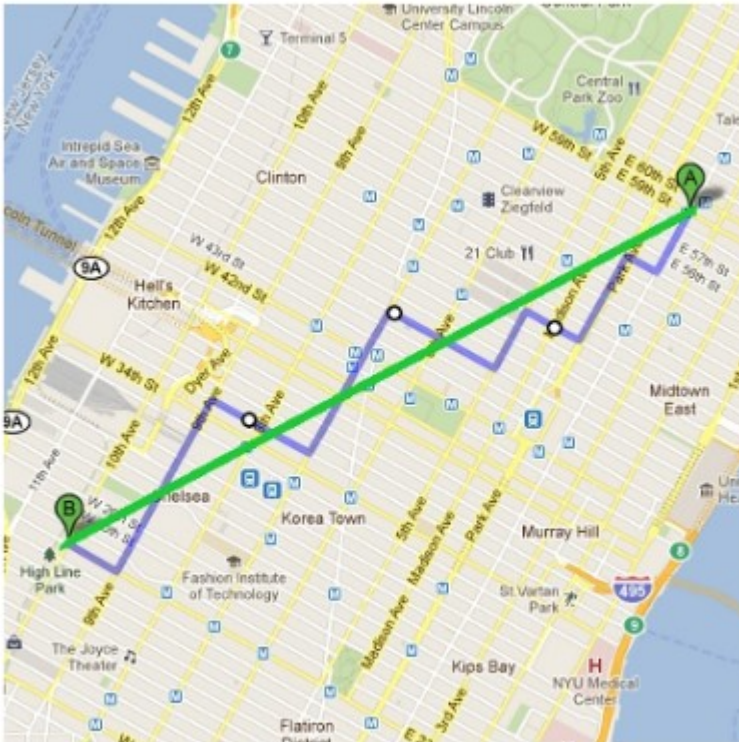
Distancias



Generalización (Minkowski):

$$Dist(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^d |x_i - y_i|^p \right)^{1/p}$$

← *Manhattan* ($p = 1$)



Distancia Manhattan

Distancias

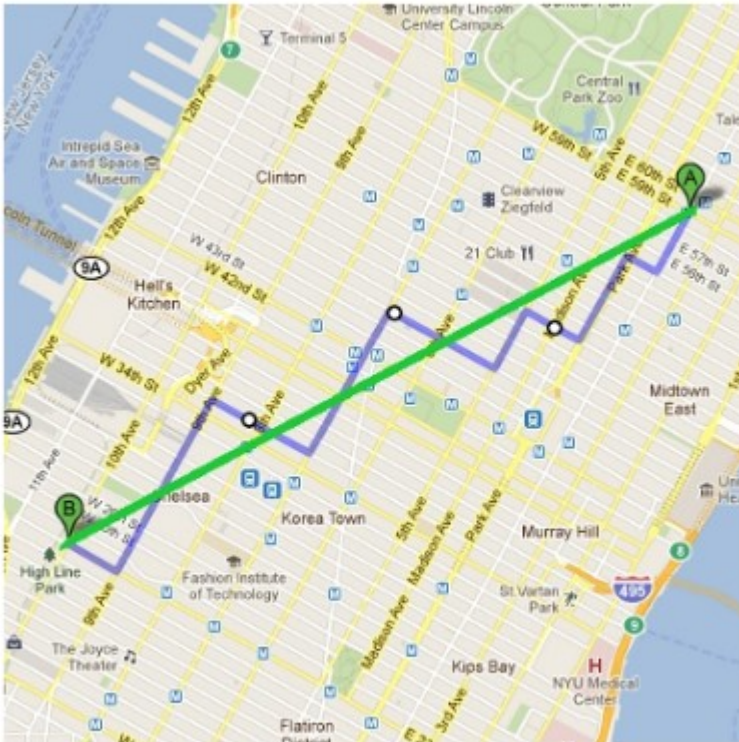


Generalización (Minkowski):

$$Dist(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^d |x_i - y_i|^p \right)^{1/p}$$

← *Manhattan* ($p = 1$)

Euclidean ($p = 2$)

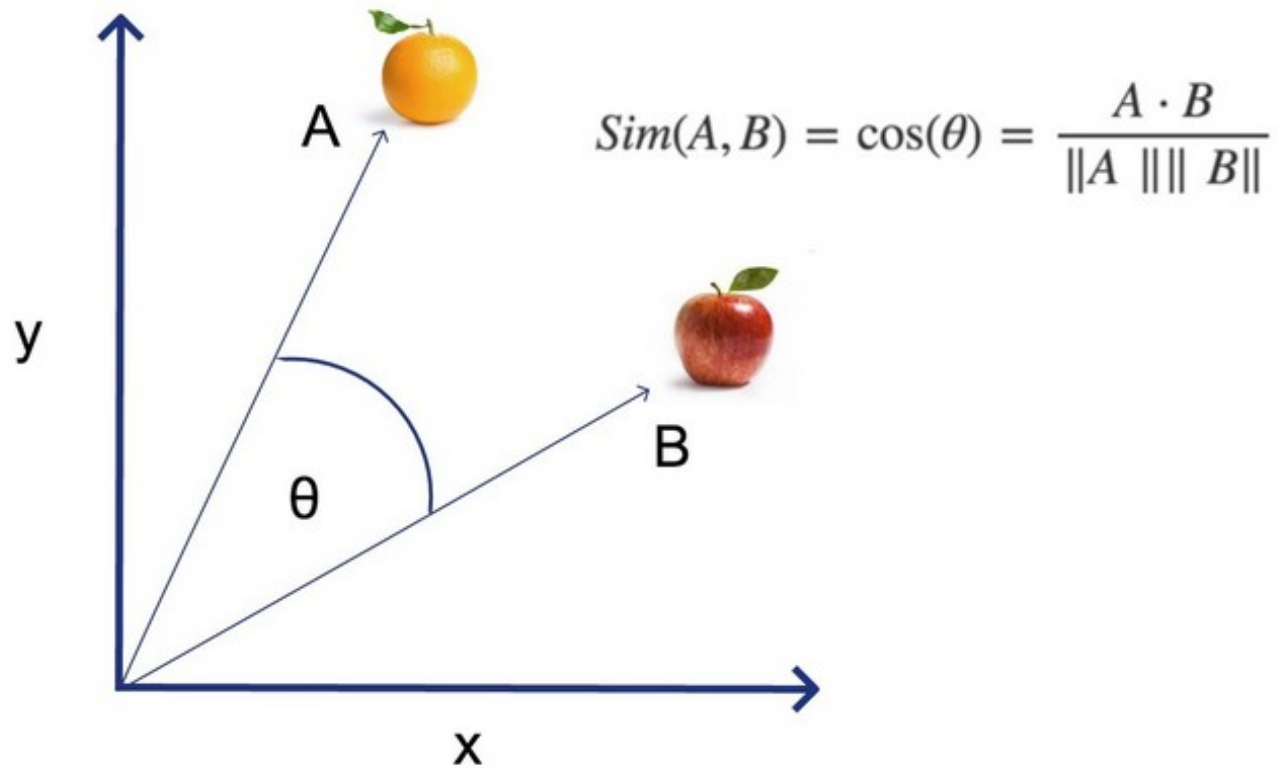


Distancia Manhattan

Proximidades

Proximidad de vectores de alta dimensionalidad:

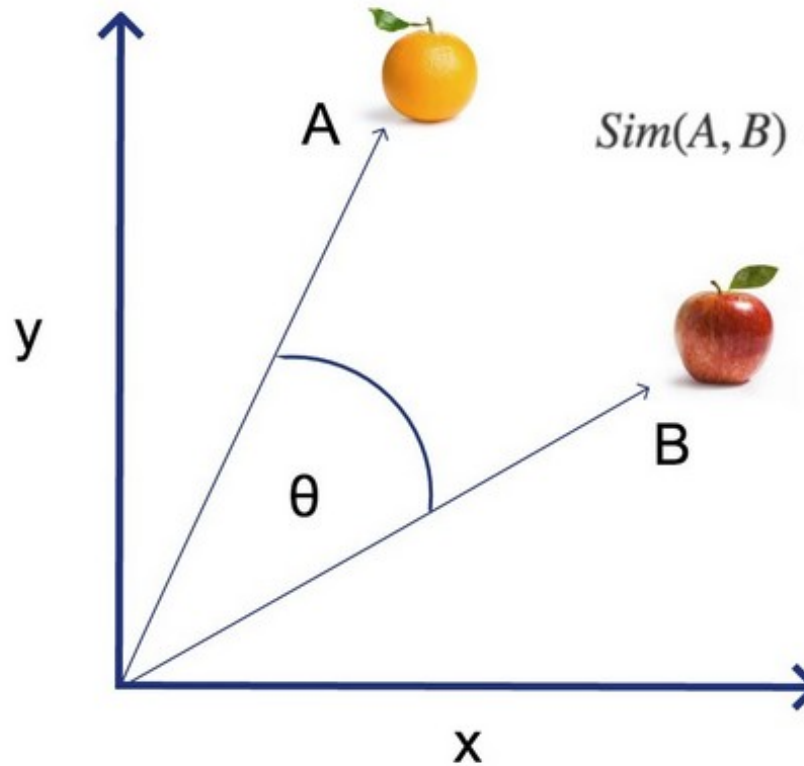
Coseno:



Proximidades

Proximidad de vectores de alta dimensionalidad:

Coseno:



$$Sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

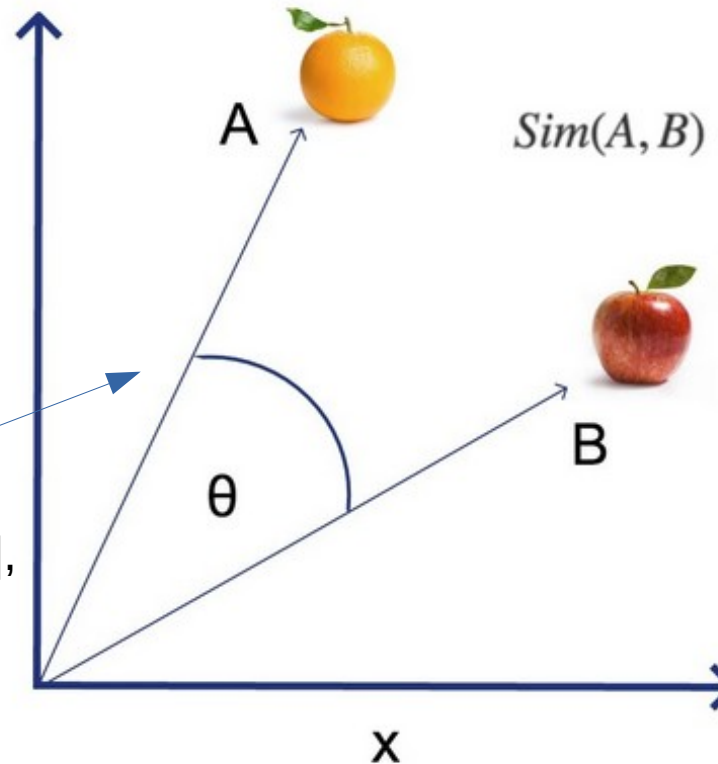
$$\cos(\bar{X}, \bar{Y}) = \frac{\sum_{i=1} x_i \cdot y_i}{\sqrt{\sum_{i=1}^d x_i^2} \cdot \sqrt{\sum_{i=1}^d y_i^2}}$$

Proximidades

Proximidad de vectores de alta dimensionalidad:

Coseno:

Si las características están en $[0,1]$,
los ángulos están en $[0^\circ, 90^\circ]$



$$Sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

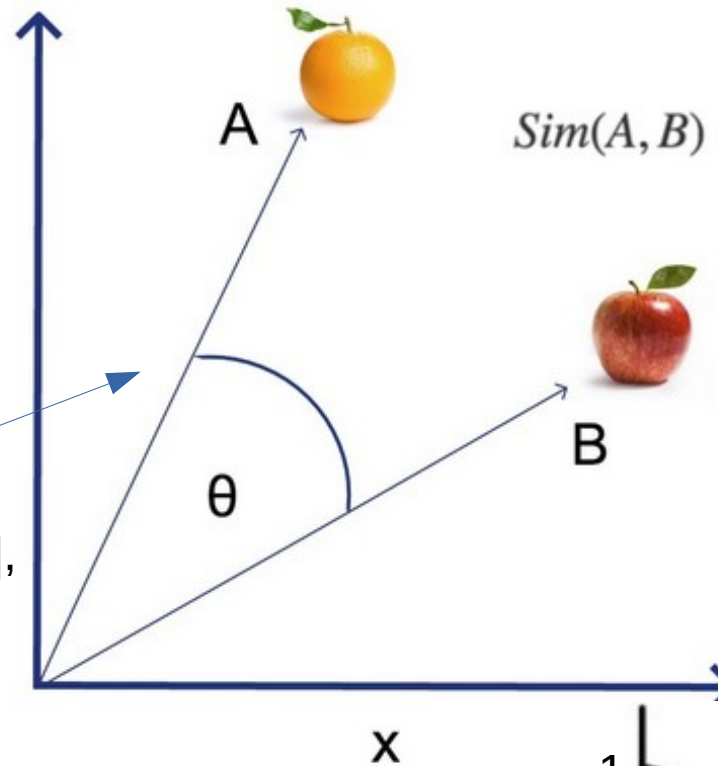
$$\cos(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^d x_i \cdot y_i}{\sqrt{\sum_{i=1}^d x_i^2} \cdot \sqrt{\sum_{i=1}^d y_i^2}}$$

Proximidades

Proximidad de vectores de alta dimensionalidad:

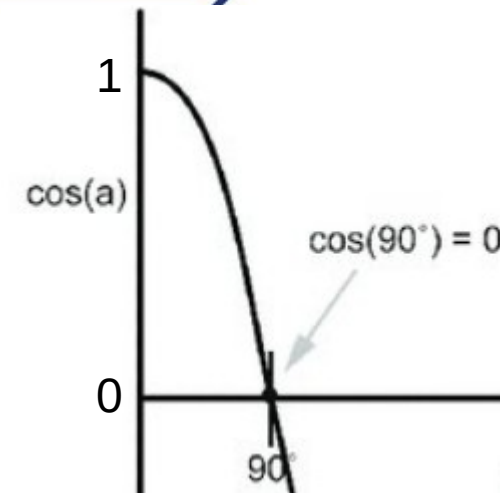
Coseno:

Si las características están en $[0,1]$,
los ángulos están en $[0^\circ, 90^\circ]$

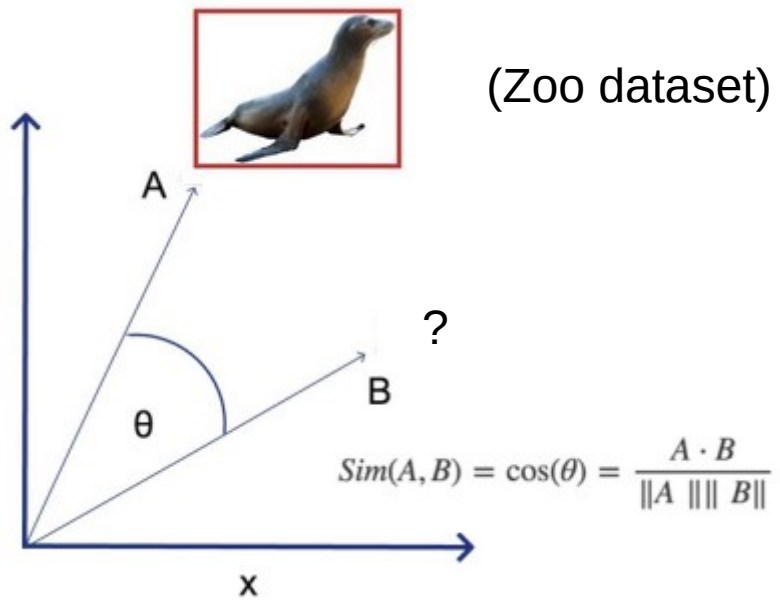


$$Sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

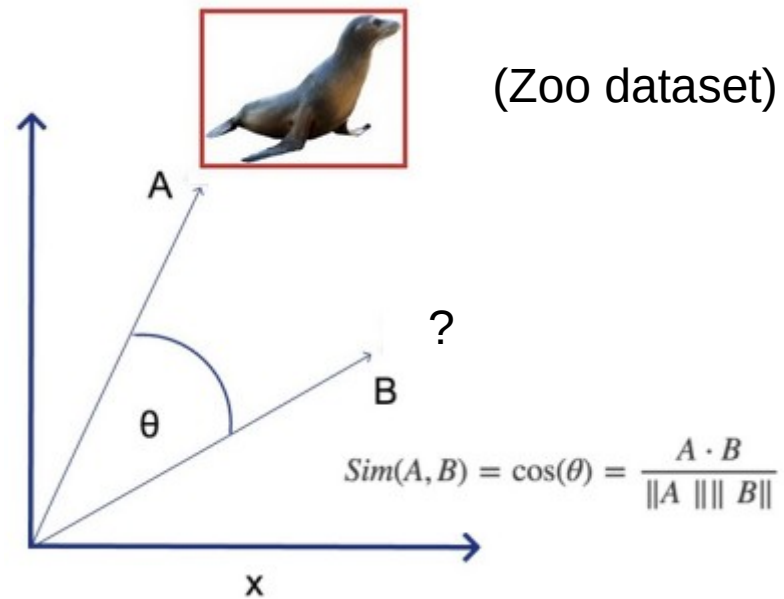
$$\cos(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^d x_i \cdot y_i}{\sqrt{\sum_{i=1}^d x_i^2} \cdot \sqrt{\sum_{i=1}^d y_i^2}}$$



Proximidades



Proximidades

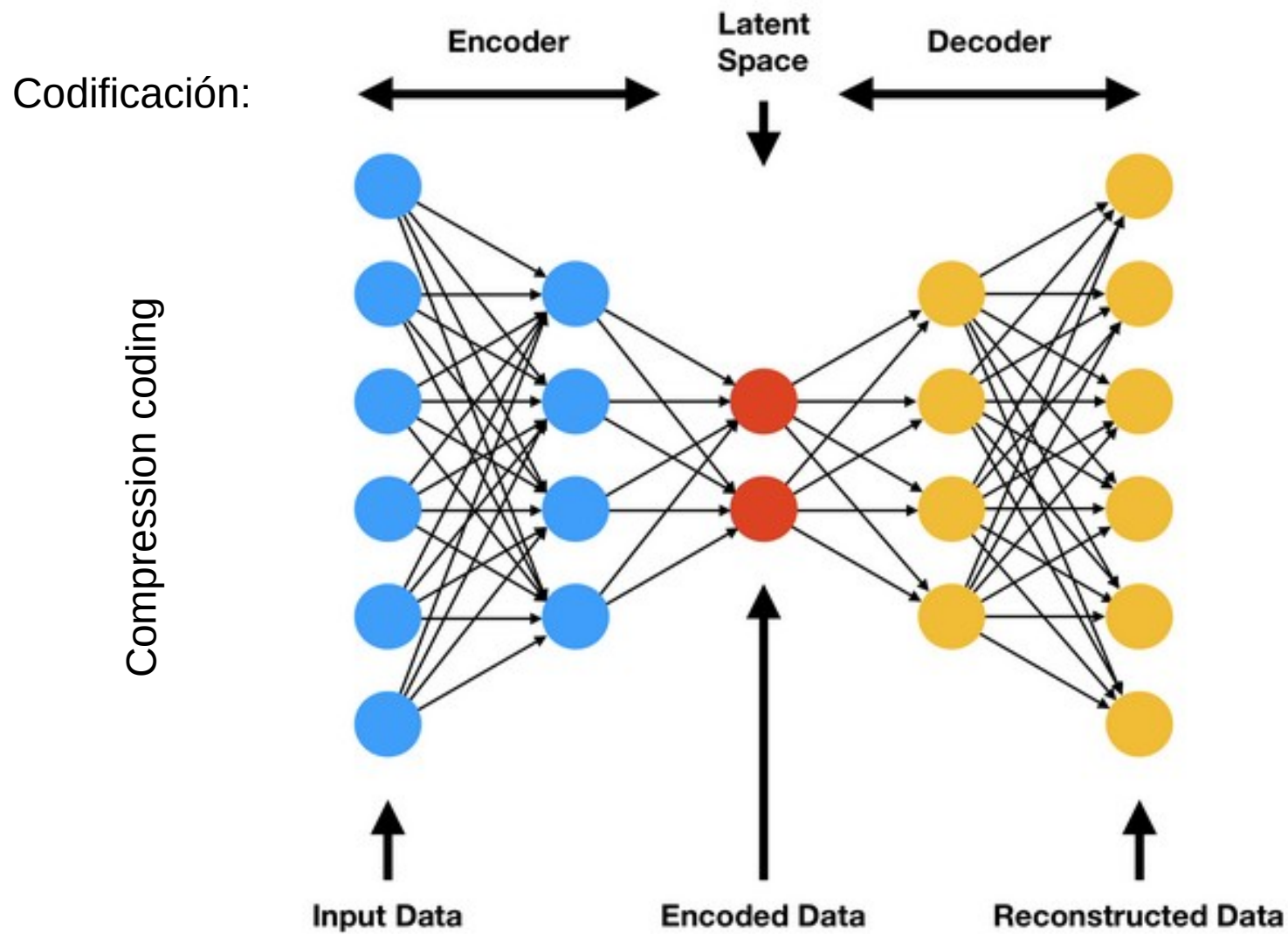


dolphin: 0.875 mink: 0.875 porpoise: 0.875 seal: 0.875 boar: 0.8125 cheetah: 0.8125 leopard: 0.8125 lion: 0.8125



- PCA -

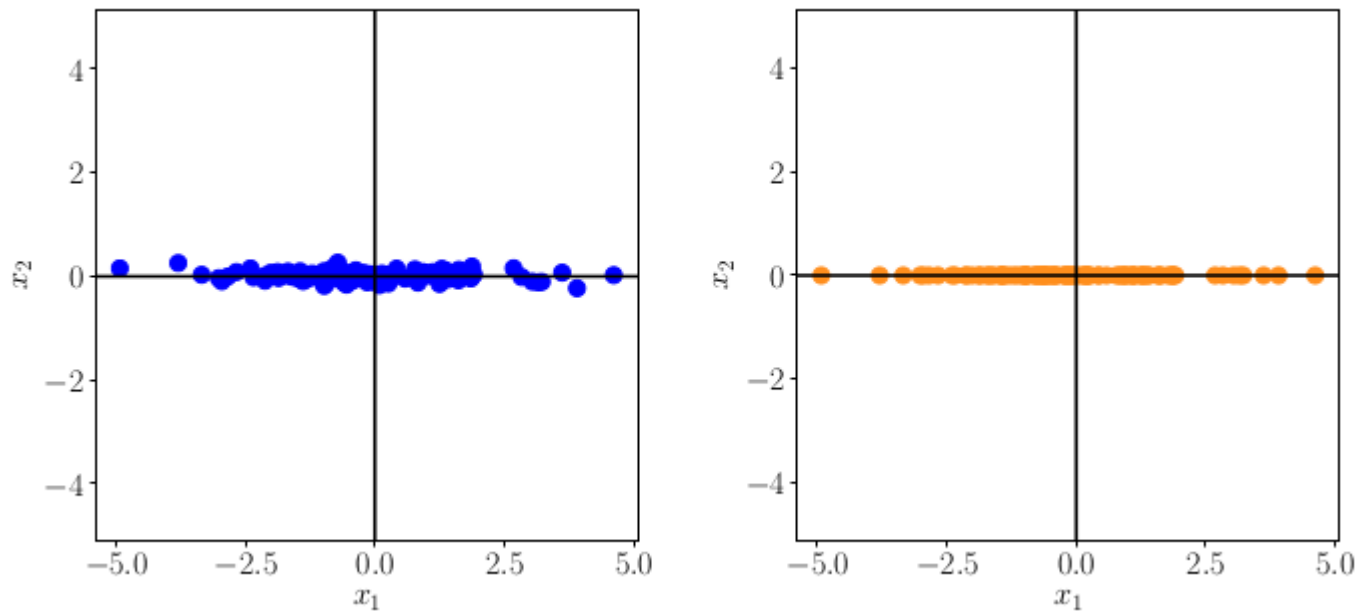
Proyección



Ej.: PCA

- UC - M. Mendoza -

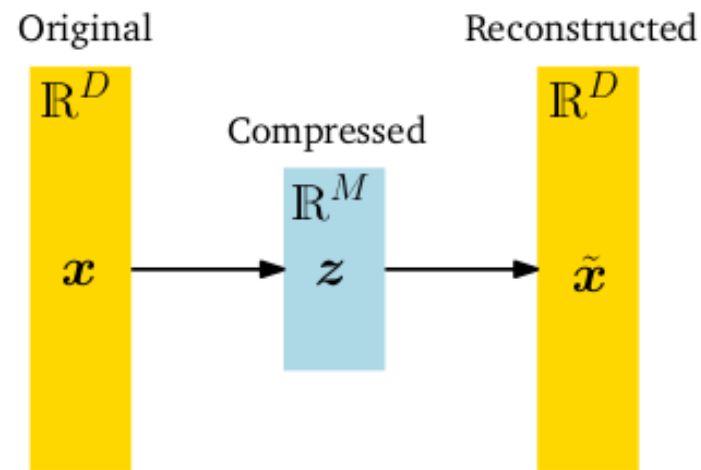
Análisis de Componentes Principales (PCA)



x_1 retiene la mayor parte de la varianza por lo que remover x_2 es neutro en términos de compresión.

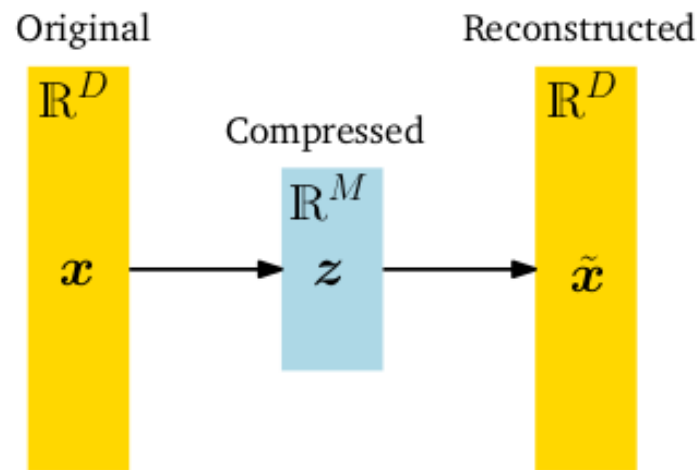
Análisis de Componentes Principales (PCA)

dataset $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, $\mathbf{x}_n \in \mathbb{R}^D$



Análisis de Componentes Principales (PCA)

dataset $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, $\mathbf{x}_n \in \mathbb{R}^D$

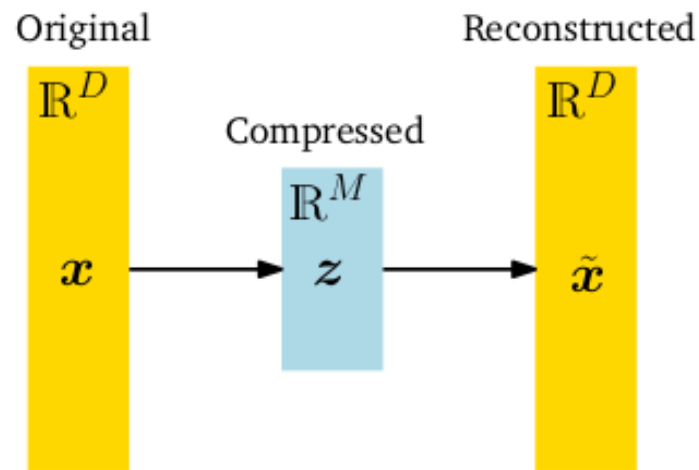


$$\mathbf{z}_n = \mathbf{B}^\top \mathbf{x}_n \in \mathbb{R}^M \longrightarrow \text{Baja dimensionalidad}$$

└─► Base de la descomposición $\mathbf{B} := [\mathbf{b}_1, \dots, \mathbf{b}_M] \in \mathbb{R}^{D \times M}$.

Análisis de Componentes Principales (PCA)

dataset $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, $\mathbf{x}_n \in \mathbb{R}^D$



$$\mathbf{z}_n = \mathbf{B}^\top \mathbf{x}_n \in \mathbb{R}^M \longrightarrow \text{Baja dimensionalidad}$$

└─► Base de la descomposición $\mathbf{B} := [\mathbf{b}_1, \dots, \mathbf{b}_M] \in \mathbb{R}^{D \times M}$.

└─► $\mathbf{b}_i^\top \mathbf{b}_j = 0$ y $\mathbf{b}_i^\top \mathbf{b}_i = 1$.

La proyección se calcula usando la SVD.

Análisis de Componentes Principales (PCA)

Proceso iterativo:

$$\hat{X} := X - \sum_{i=1}^{m-1} b_i b_i^\top X = X - B_{m-1} X, \quad \text{con } X = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$$
$$\text{y } B_{m-1} := \sum_{i=1}^{m-1} b_i b_i^\top$$

Análisis de Componentes Principales (PCA)

Proceso iterativo:

$$\hat{X} := X - \sum_{i=1}^{m-1} b_i b_i^\top X = X - B_{m-1} X, \quad \text{con } X = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$$

$$\text{y } B_{m-1} := \sum_{i=1}^{m-1} b_i b_i^\top$$

