IIC 2433 Minería de Datos

https://github.com/marcelomendoza/IIC2433

- OUTLINE -

¿Qué vamos a ver?

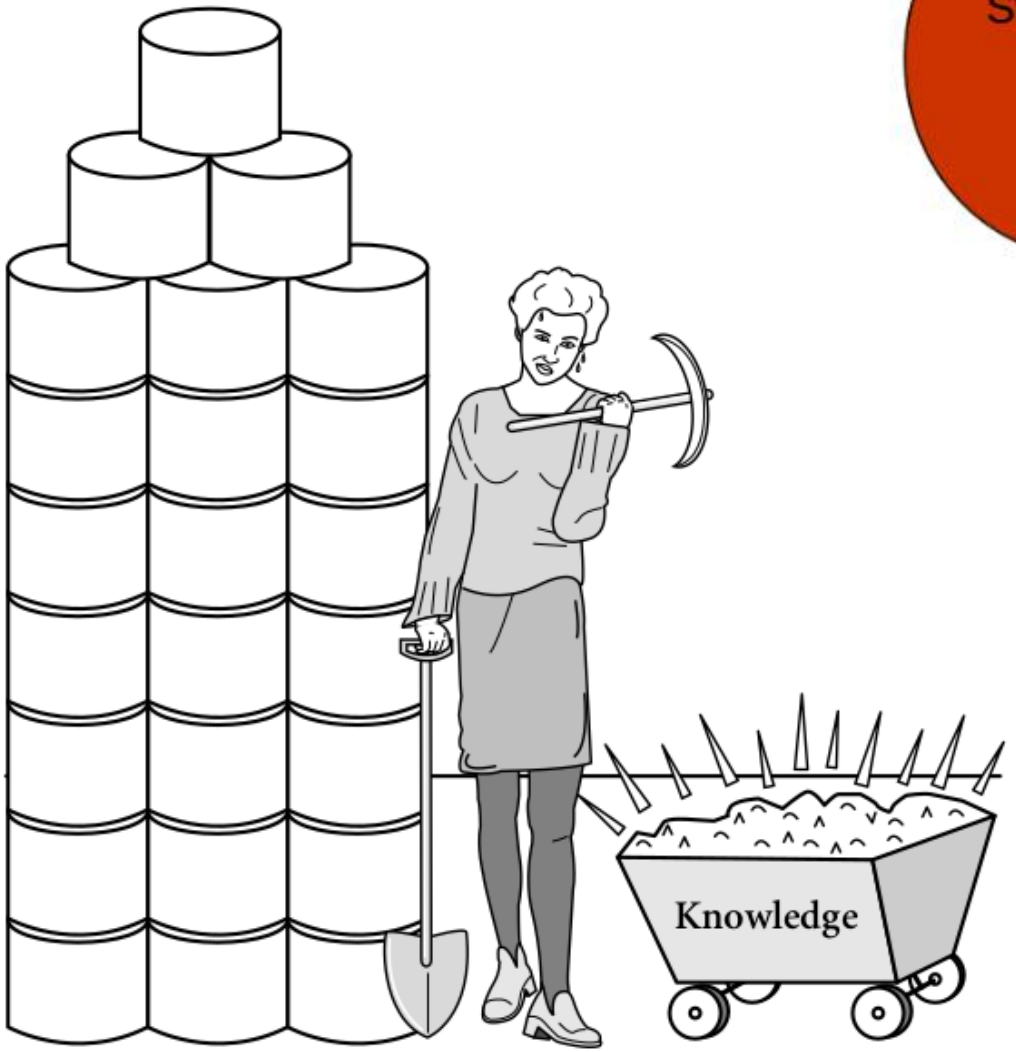
PCA, t-SNE, UMAP, outliers

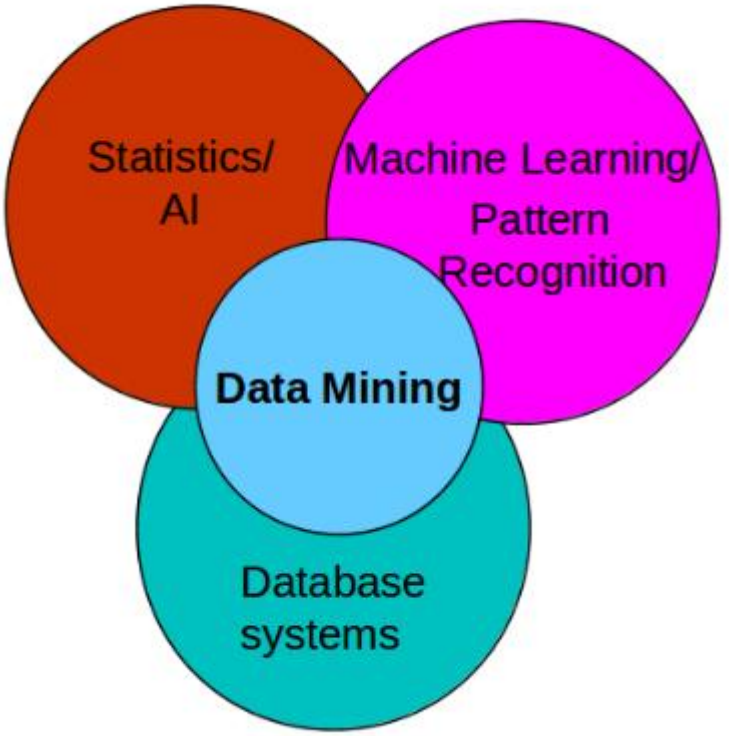
HAC, HDBSCAN, GMM

Asociaciones y causalidad

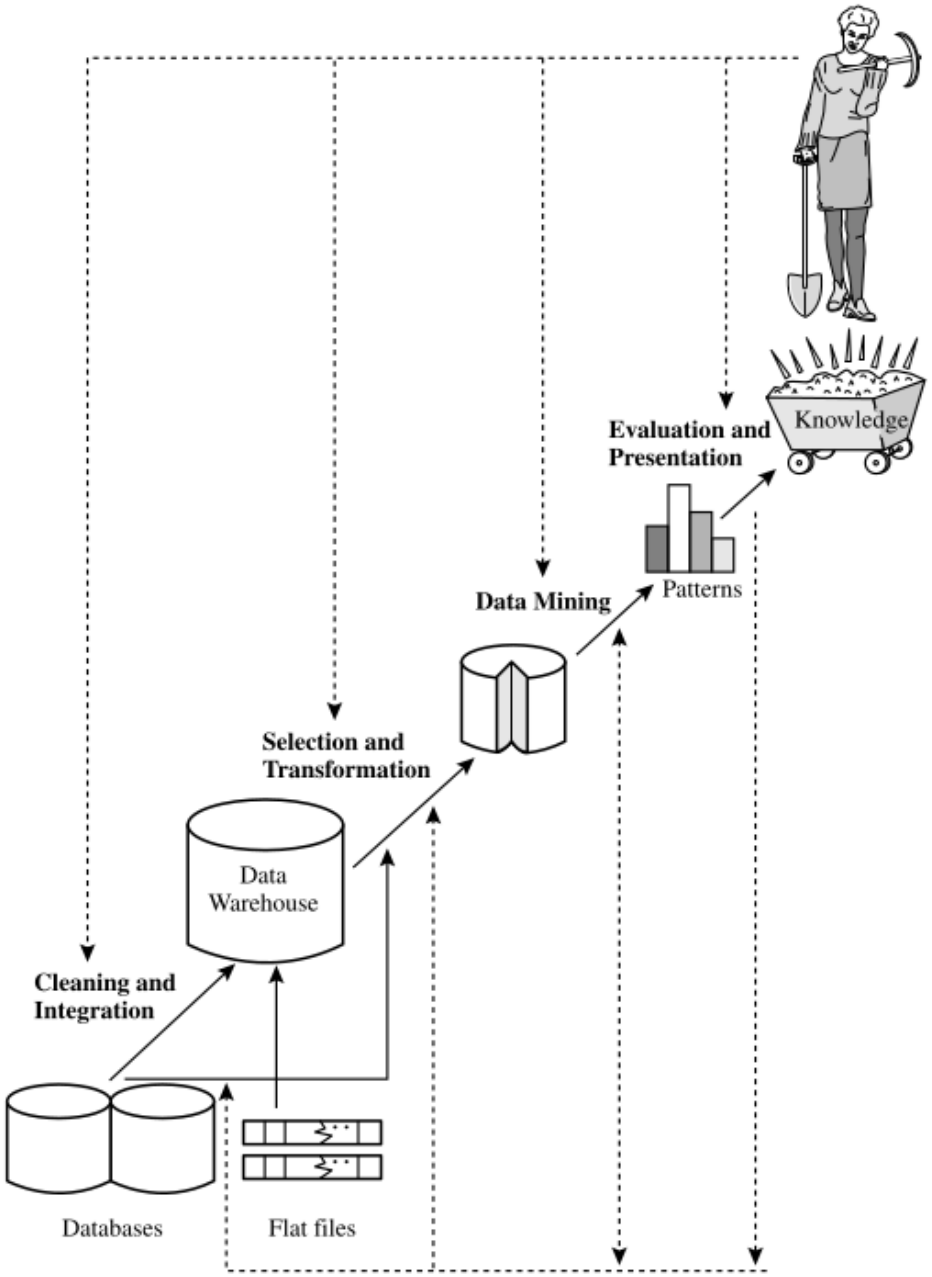
Modelos causales

Data Mining

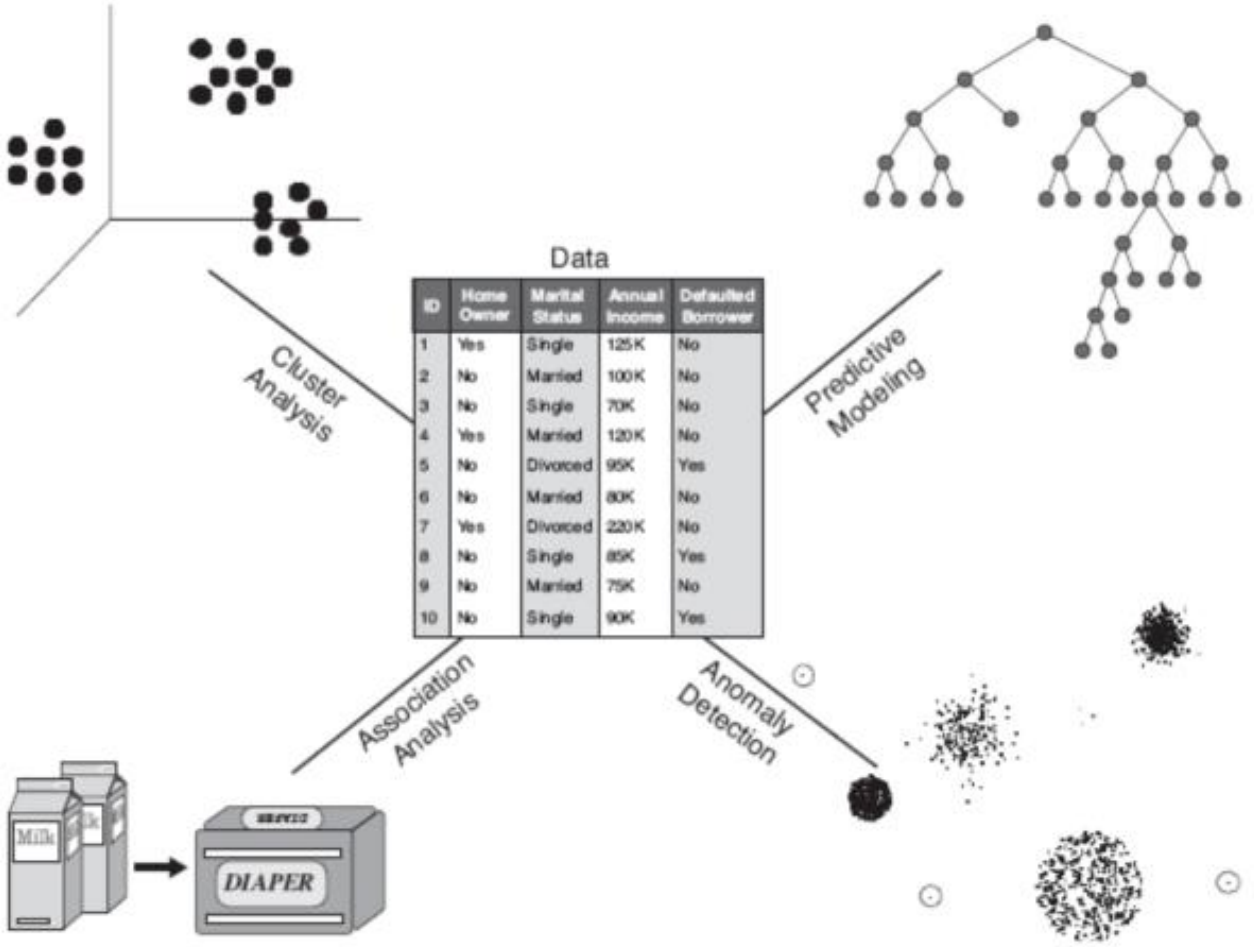

Knowledge


Statistics/
AI
Machine Learning/
Pattern
Recognition
Data Mining
Database
systems

Data Mining


Evaluation and Presentation
Knowledge
Patterns
Data Mining
Selection and Transformation
Data Warehouse
Cleaning and Integration
Databases
Flat files

Data Mining


Data
ID
Home Owner
Marital Status
Annual Income
Defaulted Borrower
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 80K No
7 Yes Divorced 220K No
8 No Single 85K Yes
9 No Married 75K No
10 No Single 90K Yes
Cluster Analysis
Predictive Modeling
Association Analysis
Anomaly Detection
DIAPER

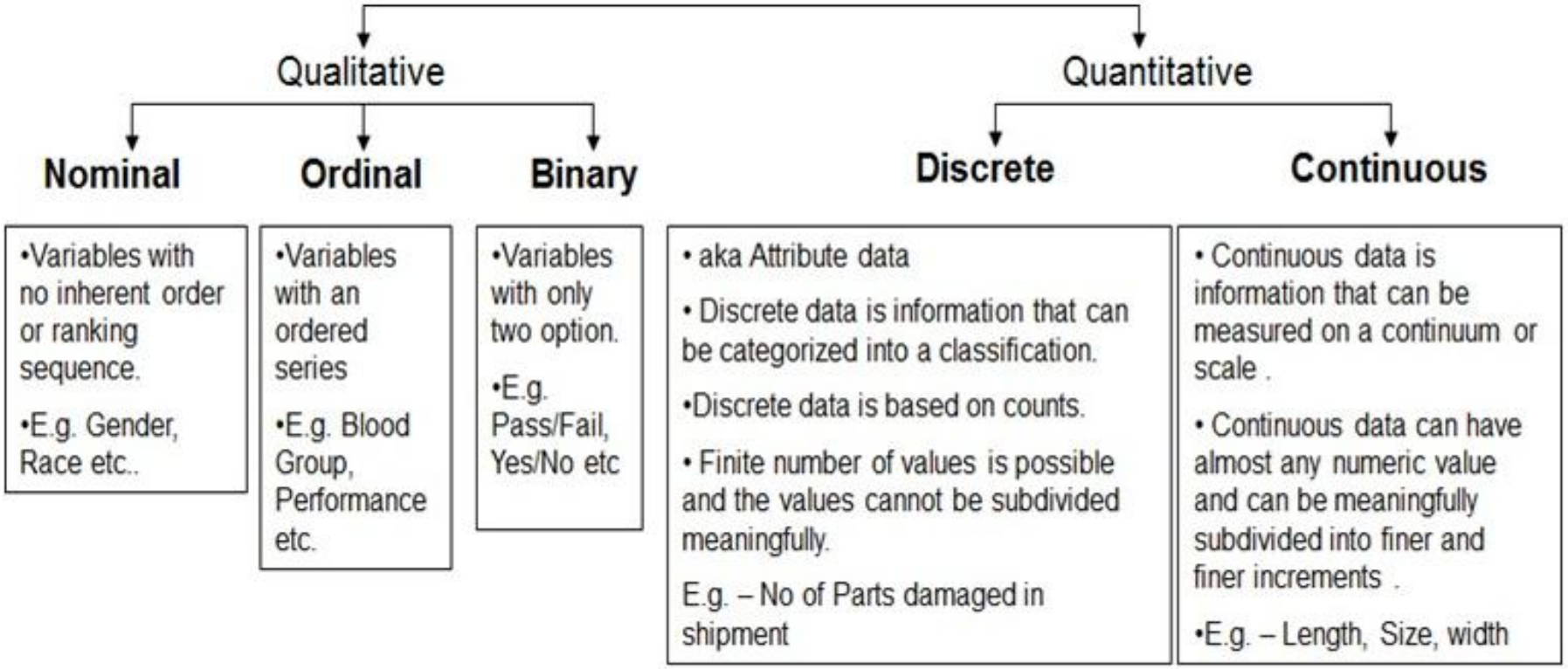
- Preprocesamiento de datos -

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:

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

Estandarización:

$$X_{m-norm} = \frac{X - \mu}{X_{max} - X_{min}}$$ ———

Intervalo centrado en torno de la media de la variable.

En general, la estandarización se centra entorno del 0.

Características cualitativas

Codificación para nominales u ordinales:

One-hot encoding

Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

La descripción

Características

Vectores y características

características

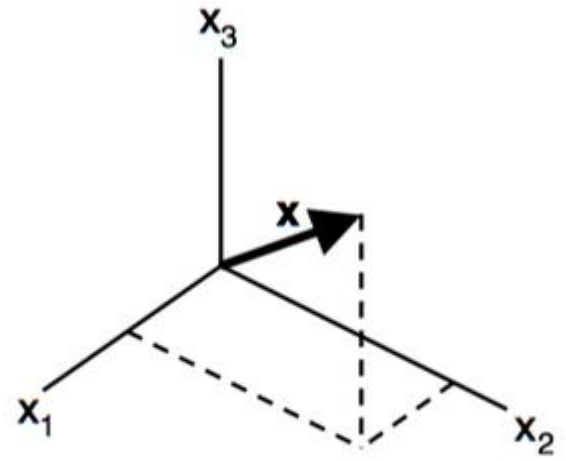
$$x = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_d \end{bmatrix}$$

Feature vector

Vectores y características

características

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

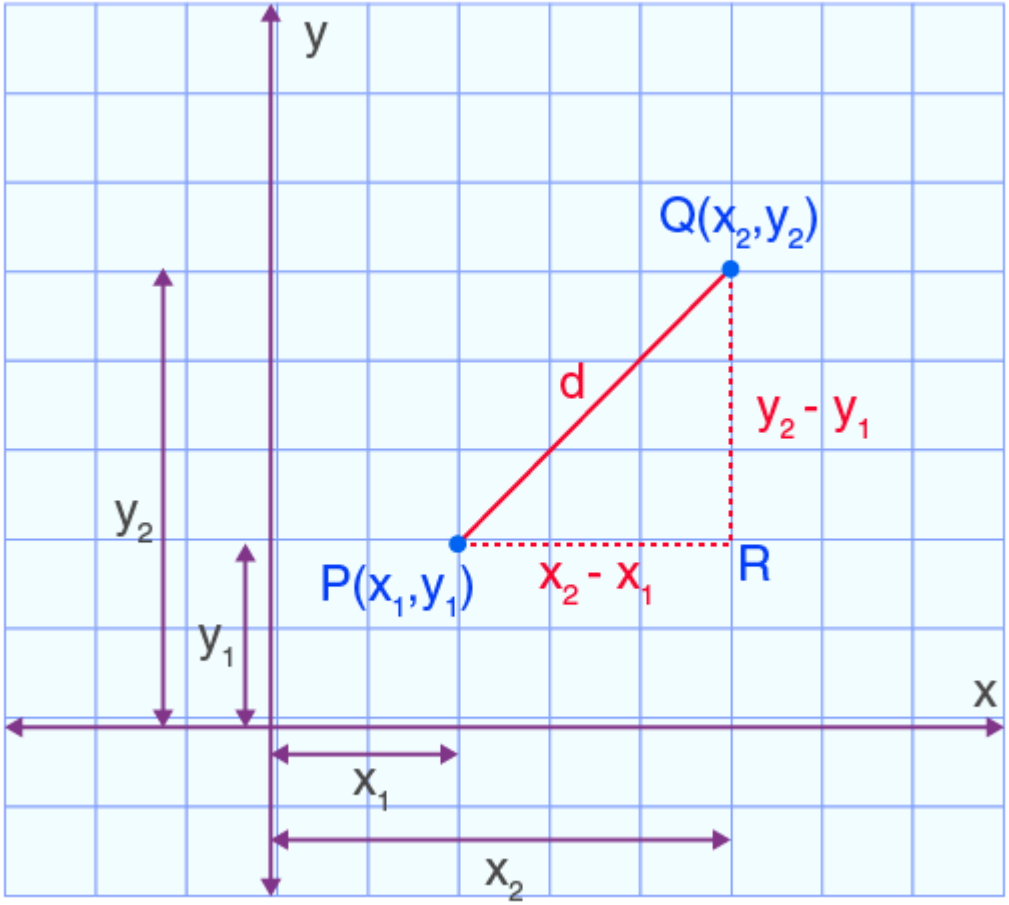


Feature vector **Feature space (3D)**

- Distancia y proximidad -

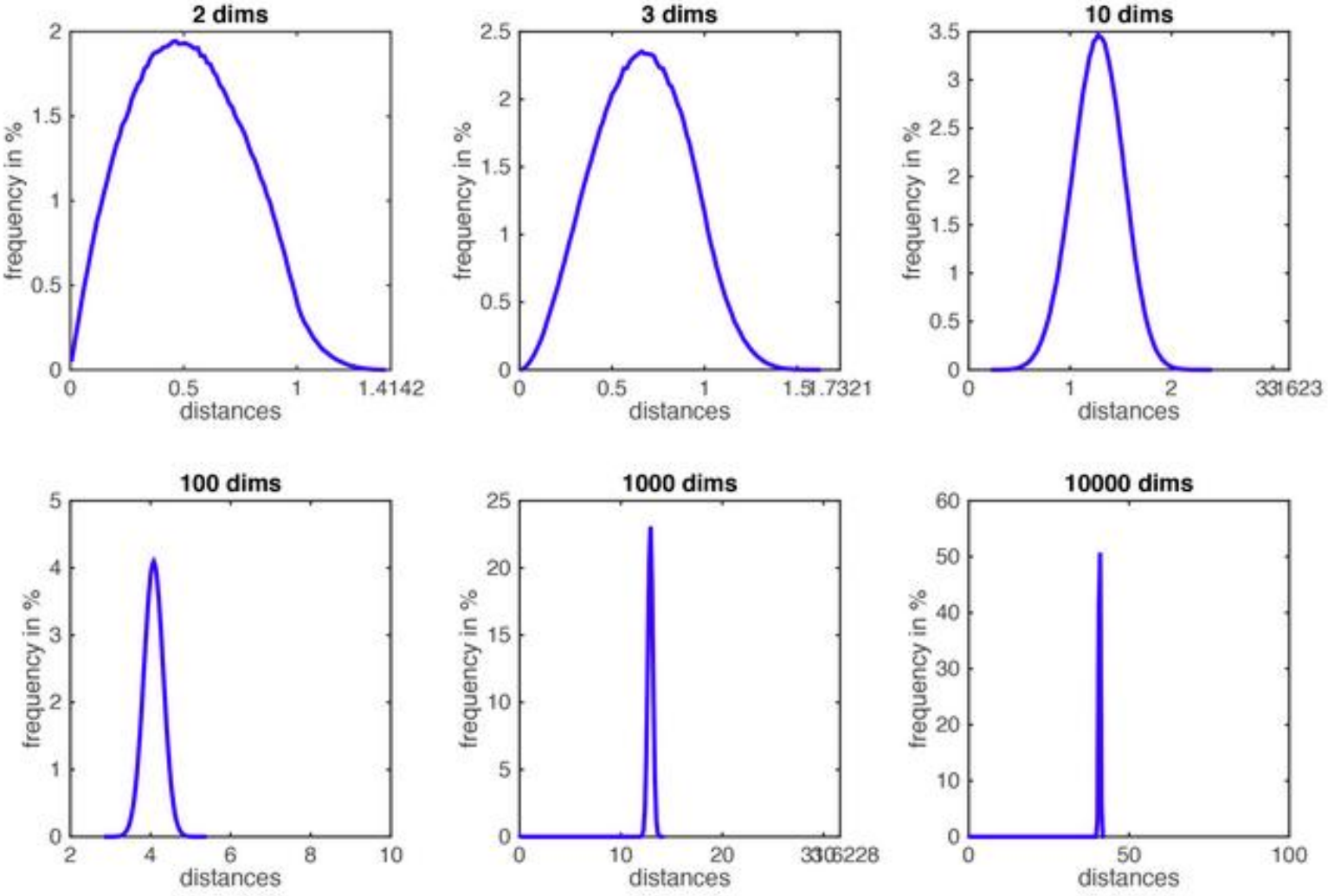
Distancia Euclideana

2D:



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

Distancia Euclideana



$\sqrt{2} = 1.41$

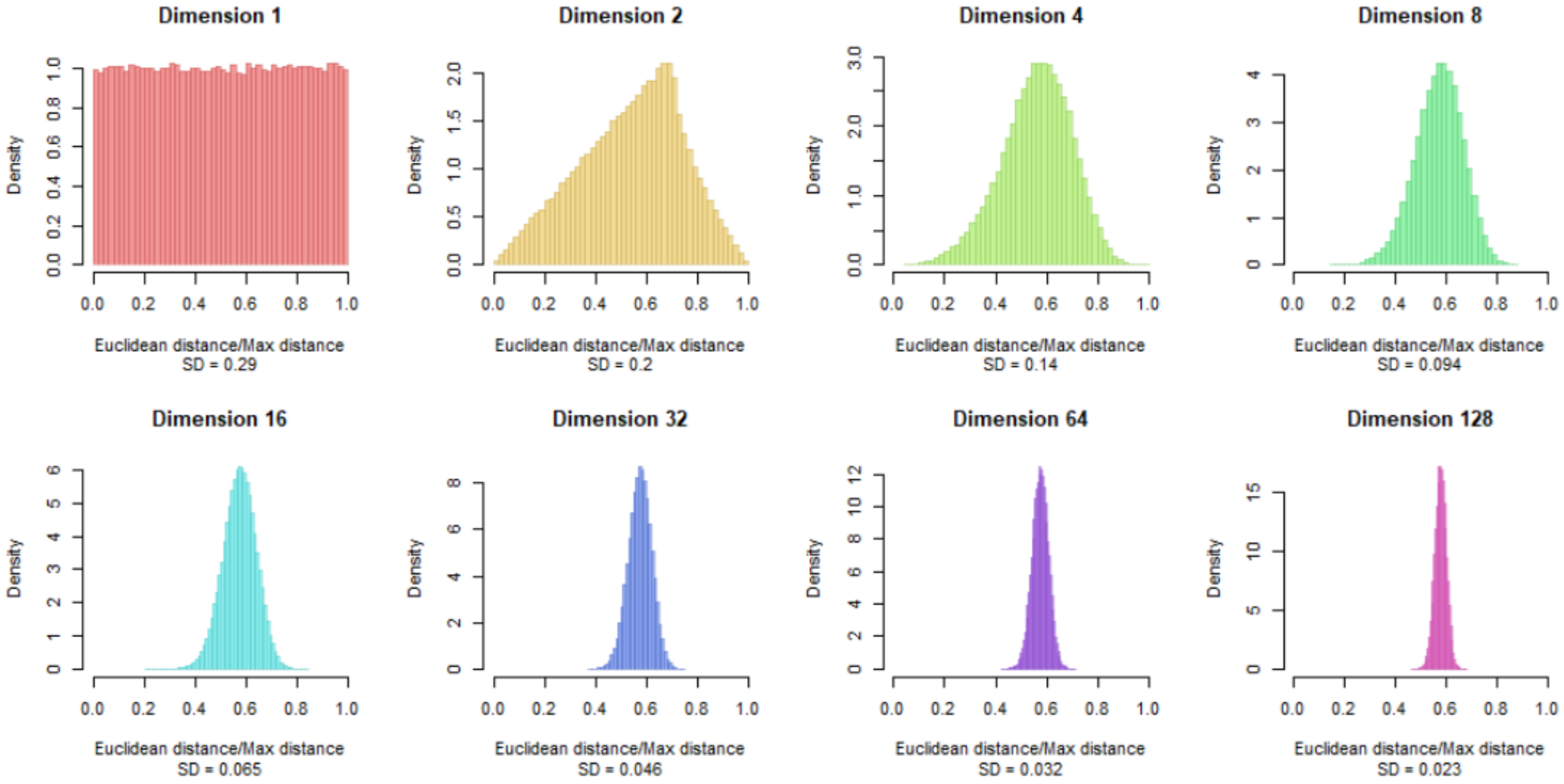
$\sqrt{3} = 1.73$

$\sqrt{10} = 3.16$

$\sqrt{1000} = 31.6$

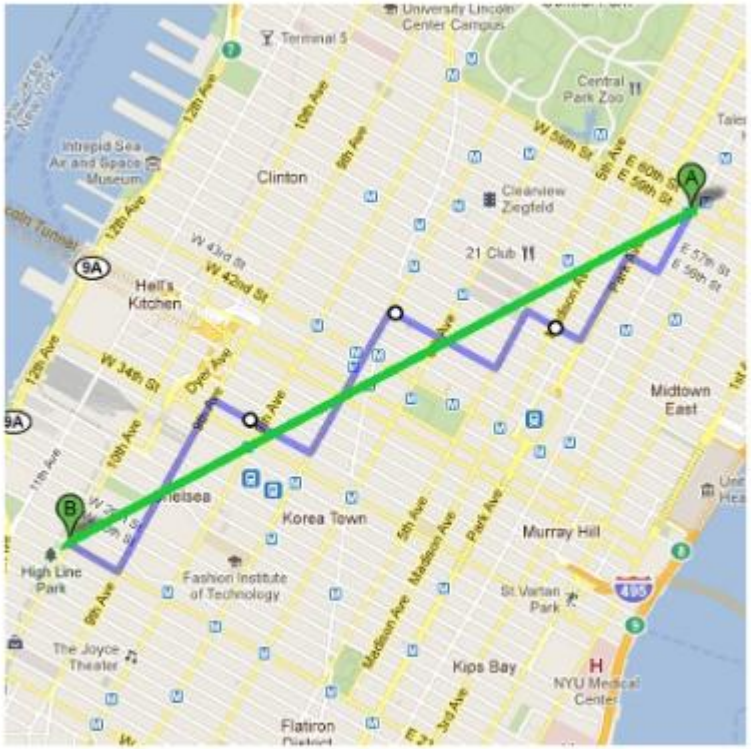
Maldición de la dimensionalidad

Distancia Euclideana



Maldición de la dimensionalidad (normalizado)

Distancias



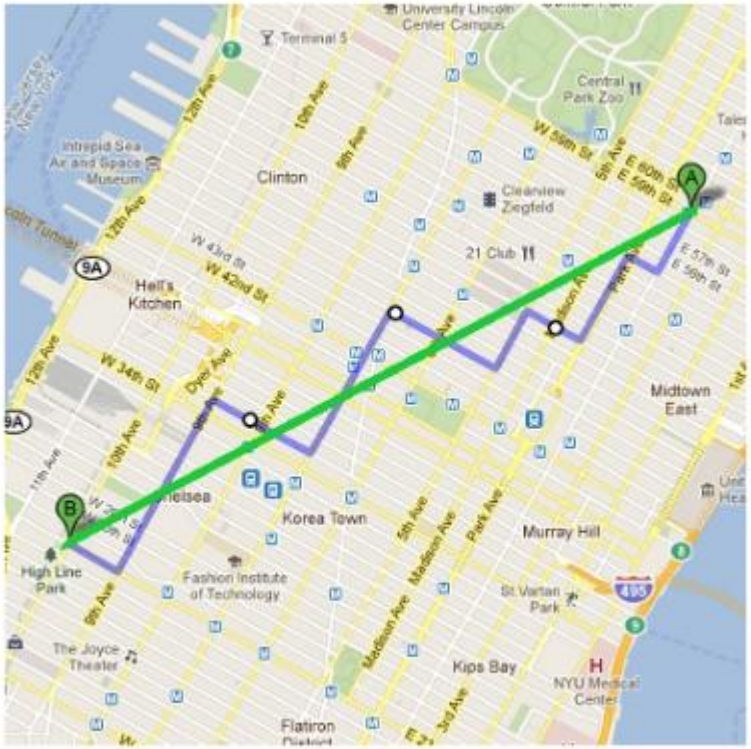
Distancia Manhattan

Distancias

Generalización (Minkowski):

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

——— $Manhattan\ (p = 1)$



Distancia Manhattan

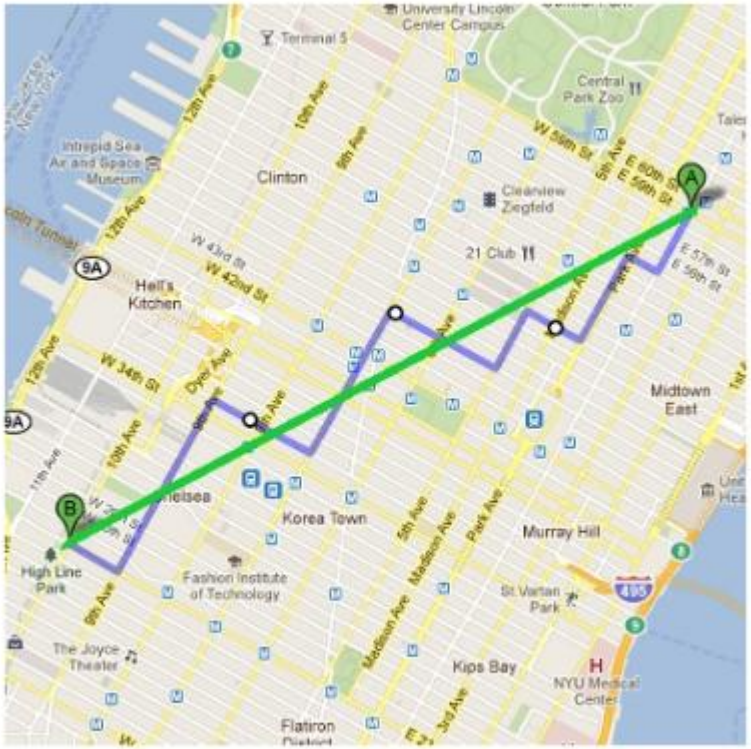
Distancias

Generalización (Minkowski):

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

$Manhattan\ (p = 1)$

$Euclidean\ (p = 2)$

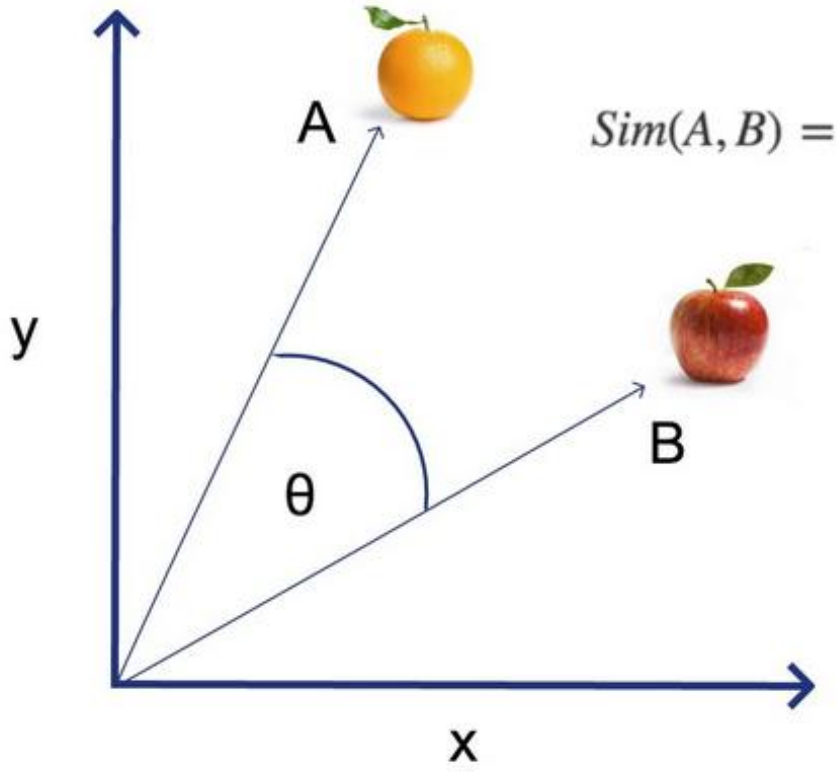


Distancia Manhattan

Proximidades

Proximidad de vectores de alta dimensionalidad:

Coseno:

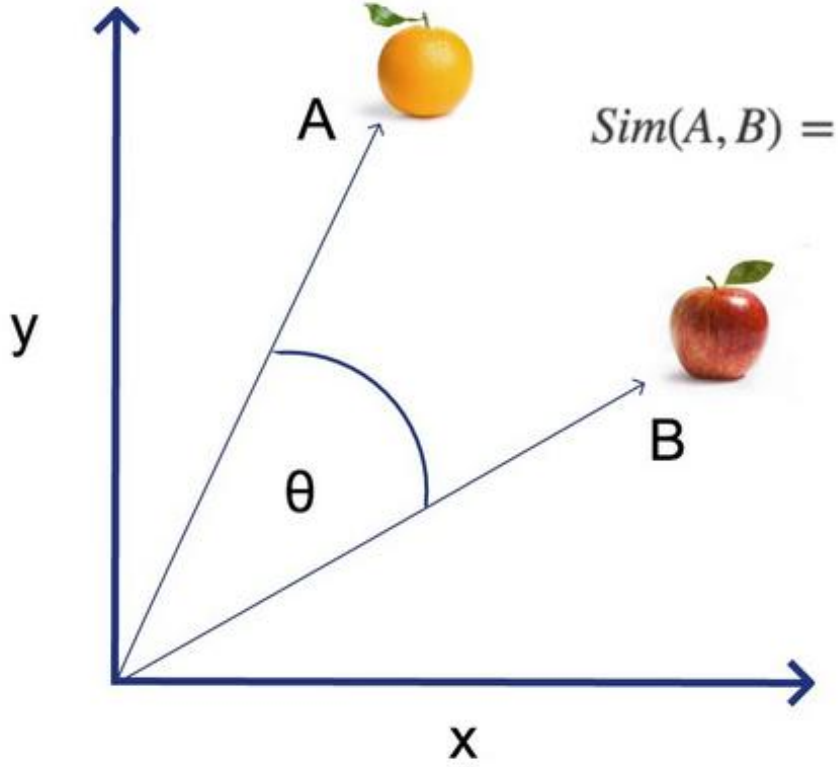


$$Sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

Proximidades

Proximidad de vectores de alta dimensionalidad:

Coseno:



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

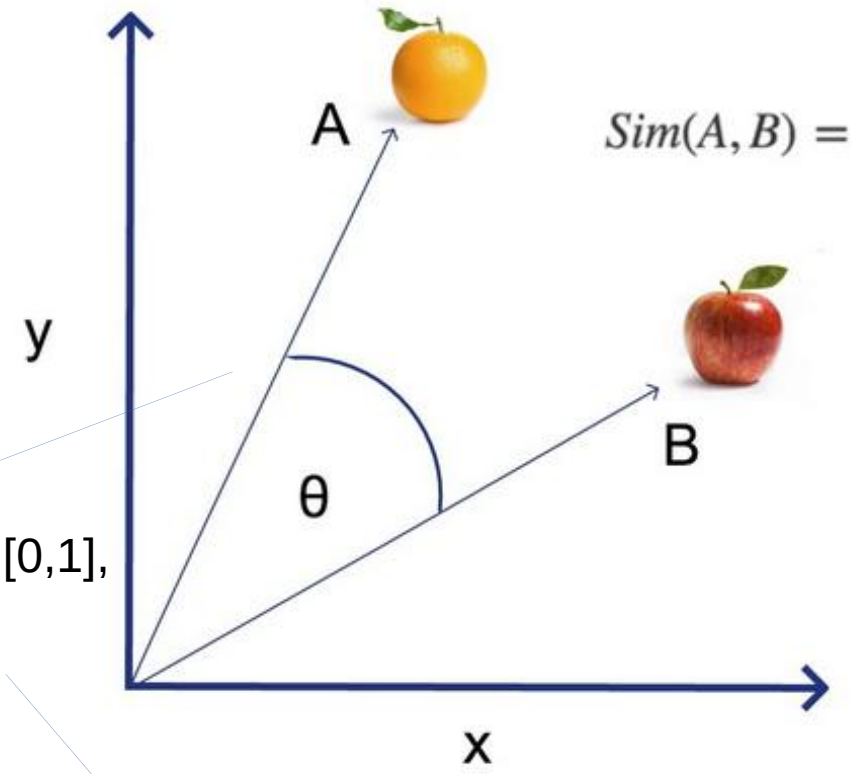
$$\cos(\overline{X}, \overline{Y}) = \frac{\sum_{i=1} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{d} x_i^2} \cdot \sqrt{\sum_{i=1}^{d} y_i^2}}$$

Proximidades

Proximidad de vectores de alta dimensionalidad:

Coseno:

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



Si las características están en [0,1], los ángulos están en [0°, 90°]

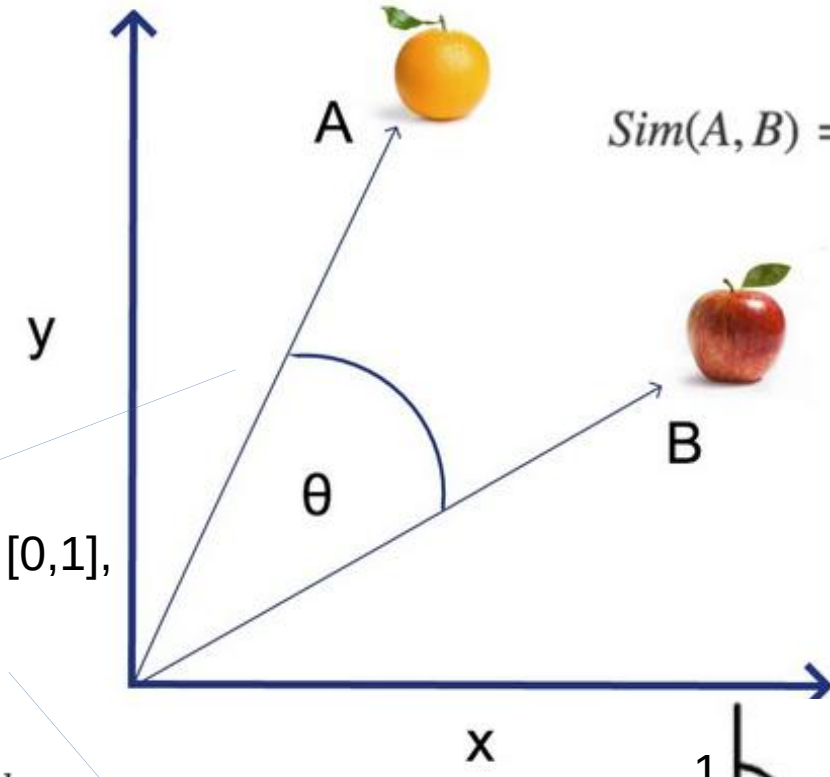
$$\cos(\overline{X}, \overline{Y}) = \frac{\sum_{i=1}^{d} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{d} x_i^2} \cdot \sqrt{\sum_{i=1}^{d} y_i^2}}$$

Proximidades

Proximidad de vectores de alta dimensionalidad:

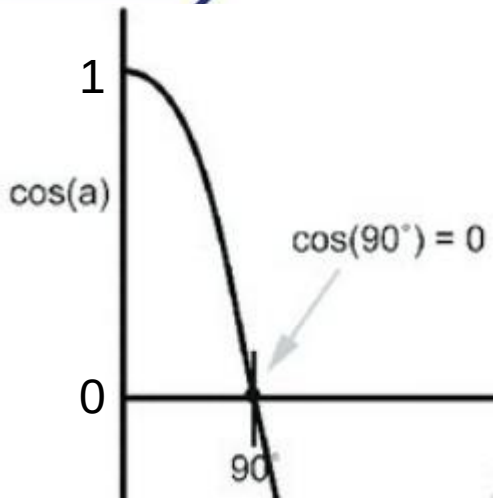
Coseno:

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



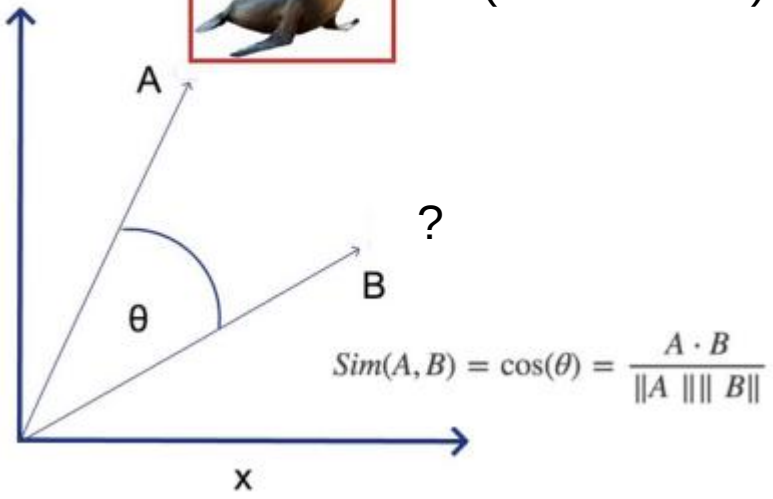
Si las características están en [0,1], los ángulos están en [0°, 90°]

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



Proximidades

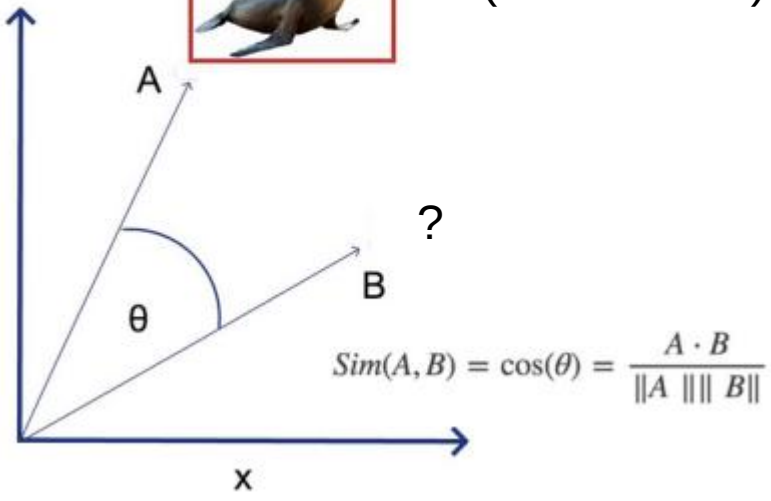
(Zoo dataset)



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

Proximidades

(Zoo dataset)



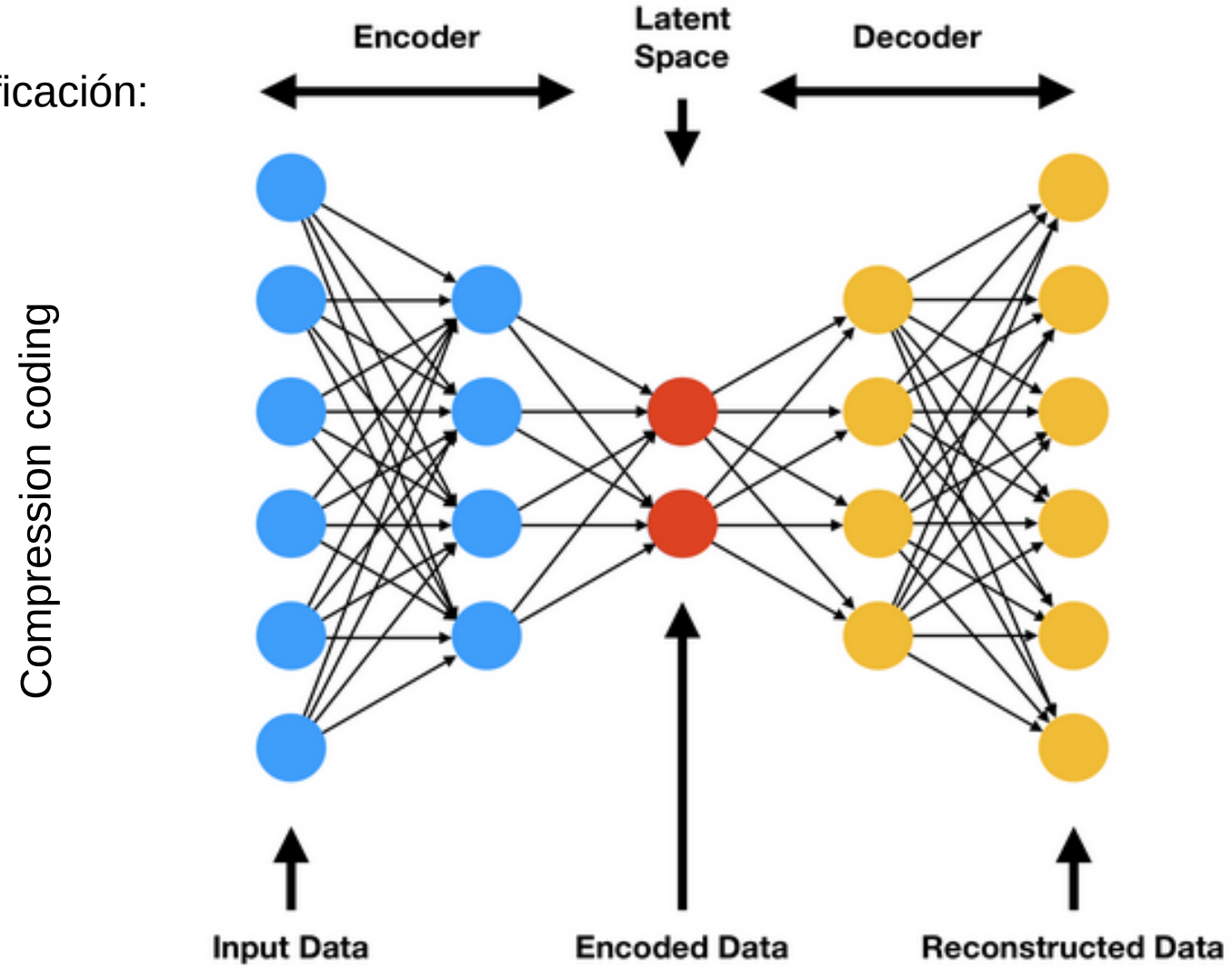
$$Sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|}$$

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 -

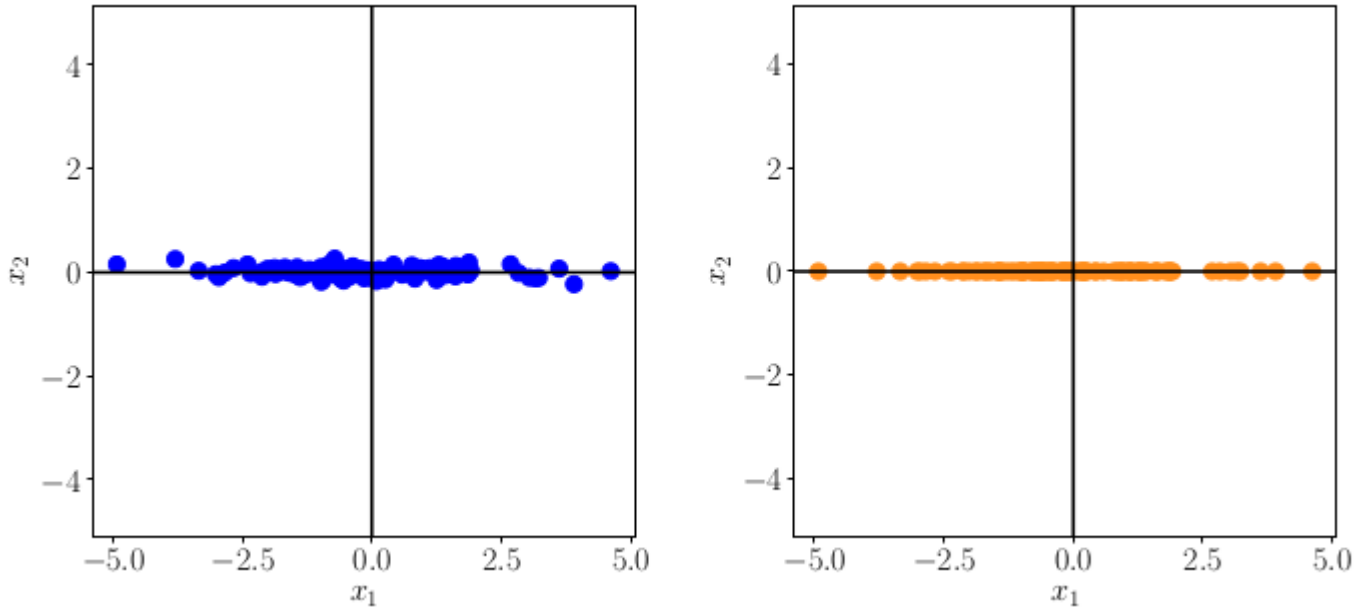
Proyección

Codificación:



Ej.: PCA

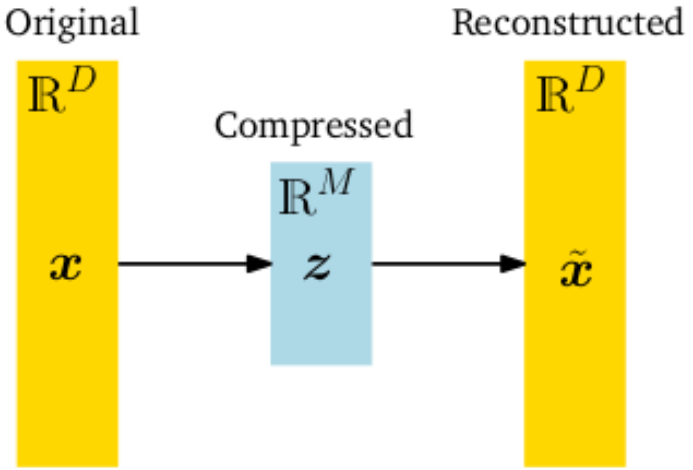
Análisis de Componentes Principales (PCA)



X1 retiene la mayor parte de la varianza por lo que remover x2 es neutro en términos de compresión.

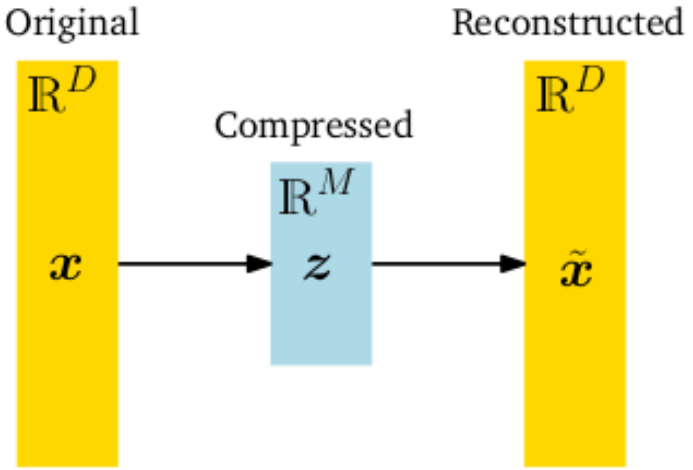
Análisis de Componentes Principales (PCA)

$$\text{dataset } \mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\},\ \boldsymbol{x}_n \in \mathbb{R}^D$$



Análisis de Componentes Principales (PCA)

$$\text{dataset } \mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\},\ \boldsymbol{x}_n \in \mathbb{R}^D$$

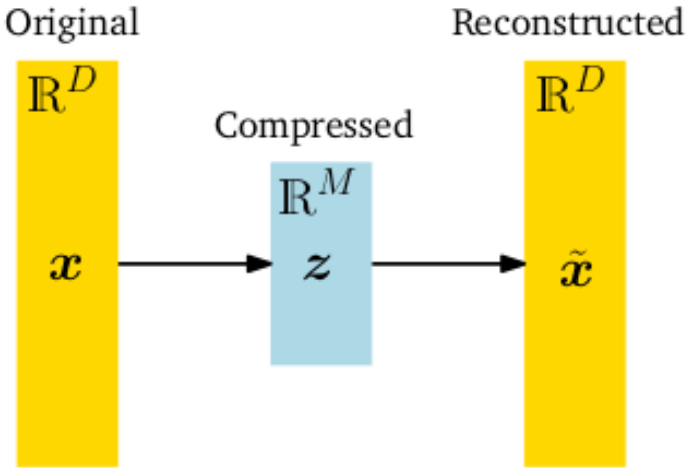


$\boldsymbol{z}_n = \boldsymbol{B}^\top \boldsymbol{x}_n \in \mathbb{R}^M$ — Baja dimensionalidad

Base de la descomposición $\boldsymbol{B} := [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M] \in \mathbb{R}^{D \times M}$.

Análisis de Componentes Principales (PCA)

$$\text{dataset } \mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\},\ \boldsymbol{x}_n \in \mathbb{R}^D$$



$$\boldsymbol{z}_n = \boldsymbol{B}^\top \boldsymbol{x}_n \in \mathbb{R}^M$$ — Baja dimensionalidad

Base ortonormal de la descomposición

$$\boldsymbol{B} := [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M] \in \mathbb{R}^{D \times M}.$$

$$\boldsymbol{b}_i^\top \boldsymbol{b}_j = 0 \quad \text{y} \quad \boldsymbol{b}_i^\top \boldsymbol{b}_i = 1$$

La proyección se calcula usando la SVD.

Análisis de Componentes Principales (PCA)

Proceso iterativo:

$$\hat{\boldsymbol{X}} := \boldsymbol{X} - \sum_{i=1}^{m-1} \boldsymbol{b}_i \boldsymbol{b}_i^\top \boldsymbol{X} = \boldsymbol{X} - \boldsymbol{B}_{m-1}\boldsymbol{X}\,, \qquad \text{con} \quad \boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D\times N}$$

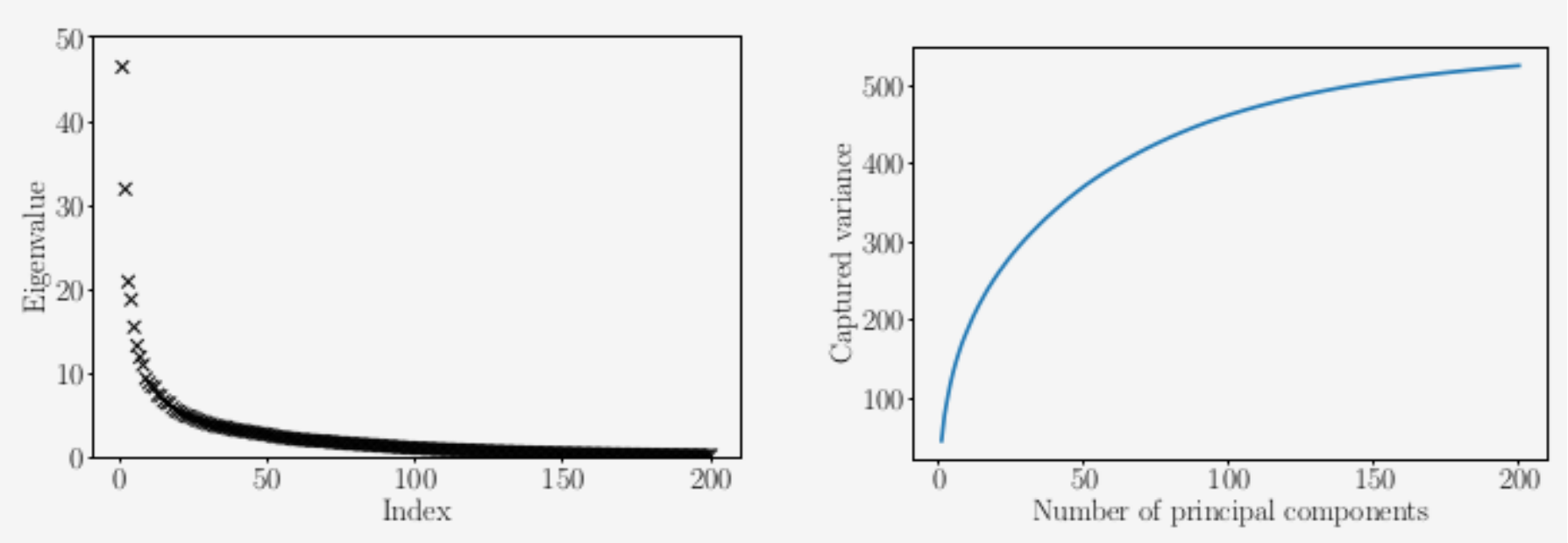
$$\text{y} \quad \boldsymbol{B}_{m-1} := \textstyle\sum_{i=1}^{m-1} \boldsymbol{b}_i \boldsymbol{b}_i^\top$$

Análisis de Componentes Principales (PCA)

Proceso iterativo:

$$\hat{\boldsymbol{X}} := \boldsymbol{X} - \sum_{i=1}^{m-1} \boldsymbol{b}_i \boldsymbol{b}_i^\top \boldsymbol{X} = \boldsymbol{X} - \boldsymbol{B}_{m-1}\boldsymbol{X}\,, \qquad \text{con} \quad \boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D\times N}$$

$$\text{y} \quad \boldsymbol{B}_{m-1} := \textstyle\sum_{i=1}^{m-1} \boldsymbol{b}_i \boldsymbol{b}_i^\top$$



Importante: puedo ajustar el # de componentes de manera que retenga una cantidad de varianza dada. Por ejemplo ¿Cuántas componentes retienen el 90% de la varianza?

Test de diagnóstico