

UMAP

Uniform Manifold Approximation and Projection

Nicolle Stefania Quintero Motta
Paula Camila García Nieto

*Minería de Datos
Pregrado en Estadística
Facultad de Ciencias - Sede Bogotá*



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Contenido

Origen
Ruta hacia UMAP
Ideas principales
Ejemplo

*Minería de Datos
Pregrado en Estadística
Facultad de Ciencias - Sede Bogotá*



UNIVERSIDAD
NACIONAL
DE COLOMBIA

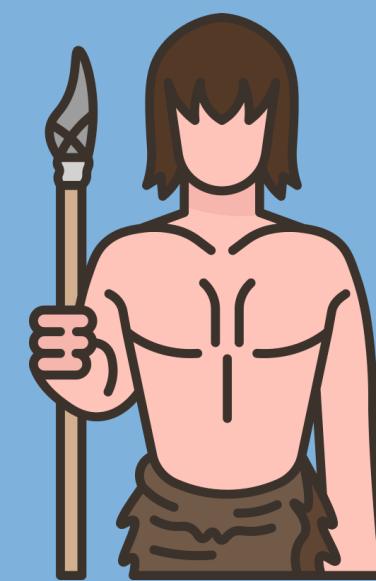


ORIGEN

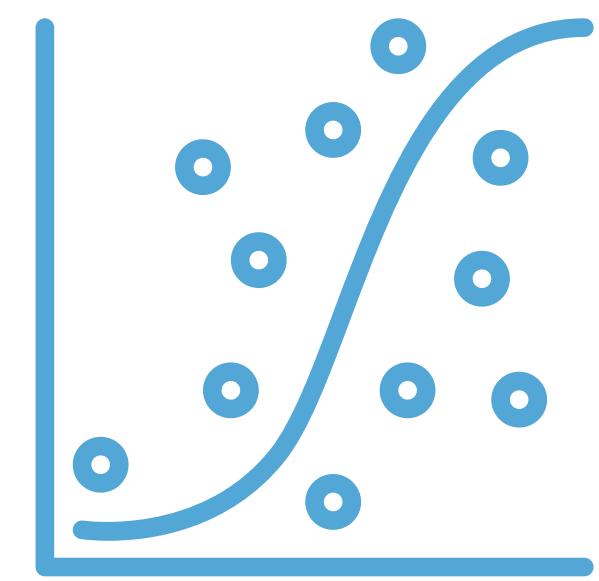
¿Para qué?



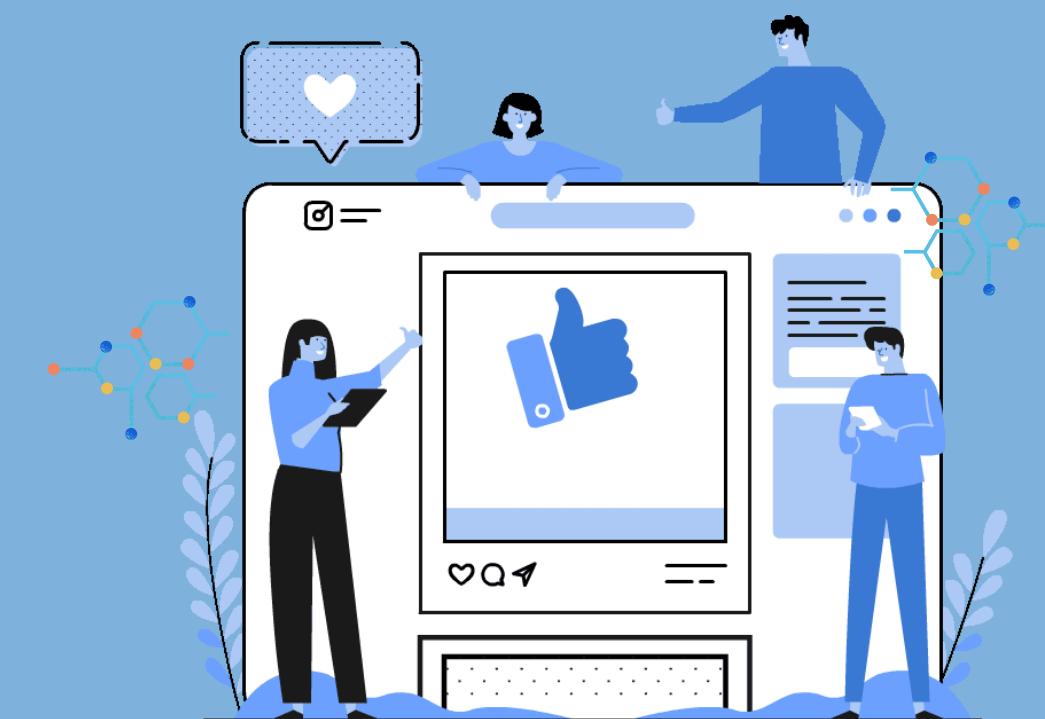
Planeación



Visualización

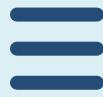


Digitalización



Reducción de la dimensionalidad

Se han desarrollado técnicas de reducción de la dimensionalidad, cada vez más eficientes para que al momento de la reducción de variables se pierda la menor cantidad de información posible.



HACIA UMAP

Uniform Manifold Approximation and Projection



Minería de Datos
Pregrado en Estadística
Facultad de Ciencias - Sede Bogotá



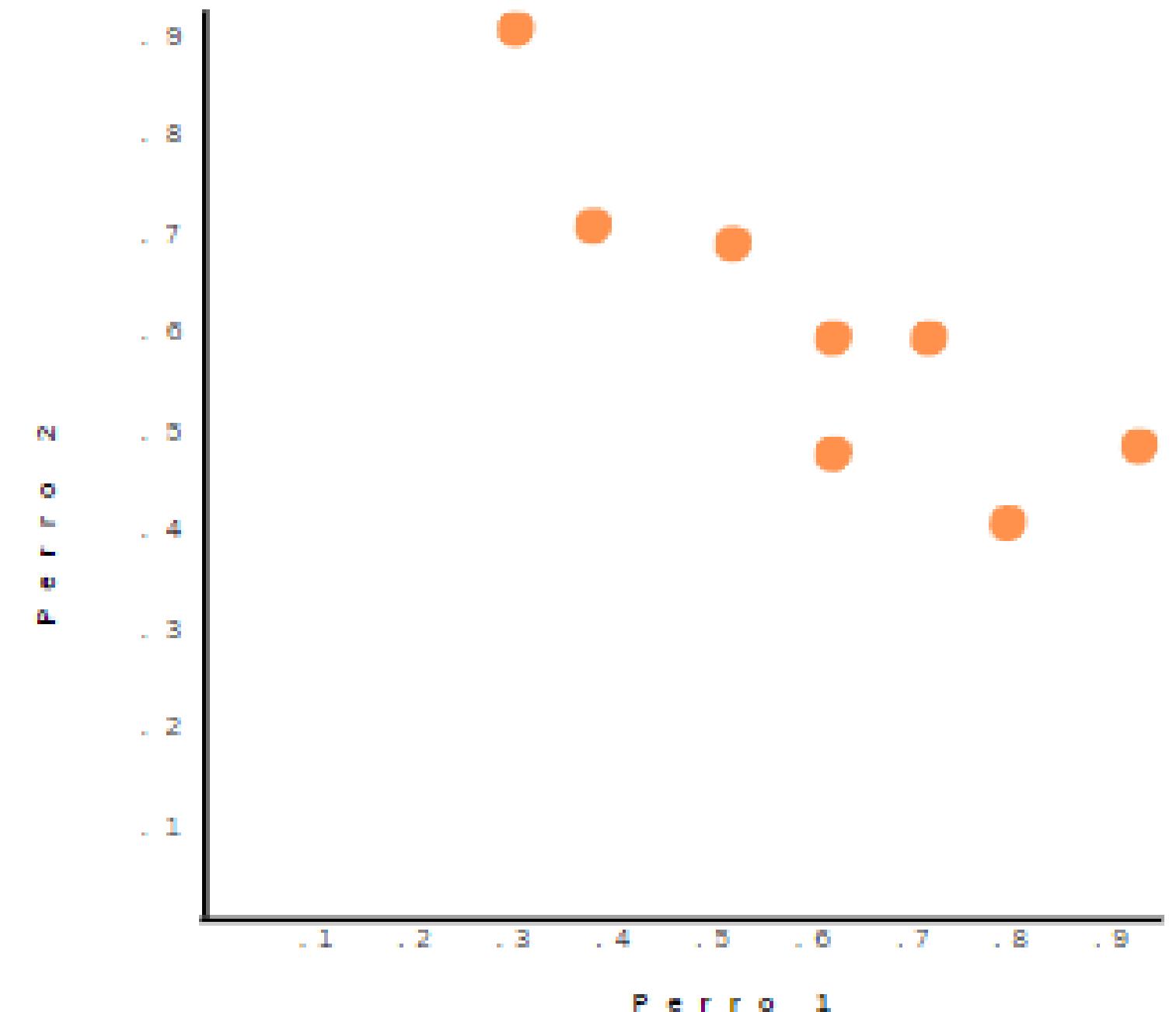
PCA (Principal Component Analysis)

Registro de raza	Perro1	Perro2	Perro3	Perro4	Perro5
Gen1	0.8	0.4	0.9	0.4	0.6
Gen2	0.4	0.7	0.3	0.3	0.5
Gen3	0.7	0.6	0.7	0.9	0.9
Gen4	0.3	0.9	0.4	0.9	0.6
Gen5	0.6	0.6	0.7	0.8	0.6
Gen6	0.5	0.7	0.5	0.6	0.8
Gen7	0.6	0.5	0.6	0.4	0.5
Gen8	0.9	0.5	0.8	0.8	0.7

PCA (Principal Component Analysis)

Registro de raza	Perro1	Perro2
Gen1	0.8	0.4
Gen2	0.4	0.7
Gen3	0.7	0.6
Gen4	0.3	0.9
Gen5	0.6	0.6
Gen6	0.5	0.7
Gen7	0.6	0.5
Gen8	0.9	0.5

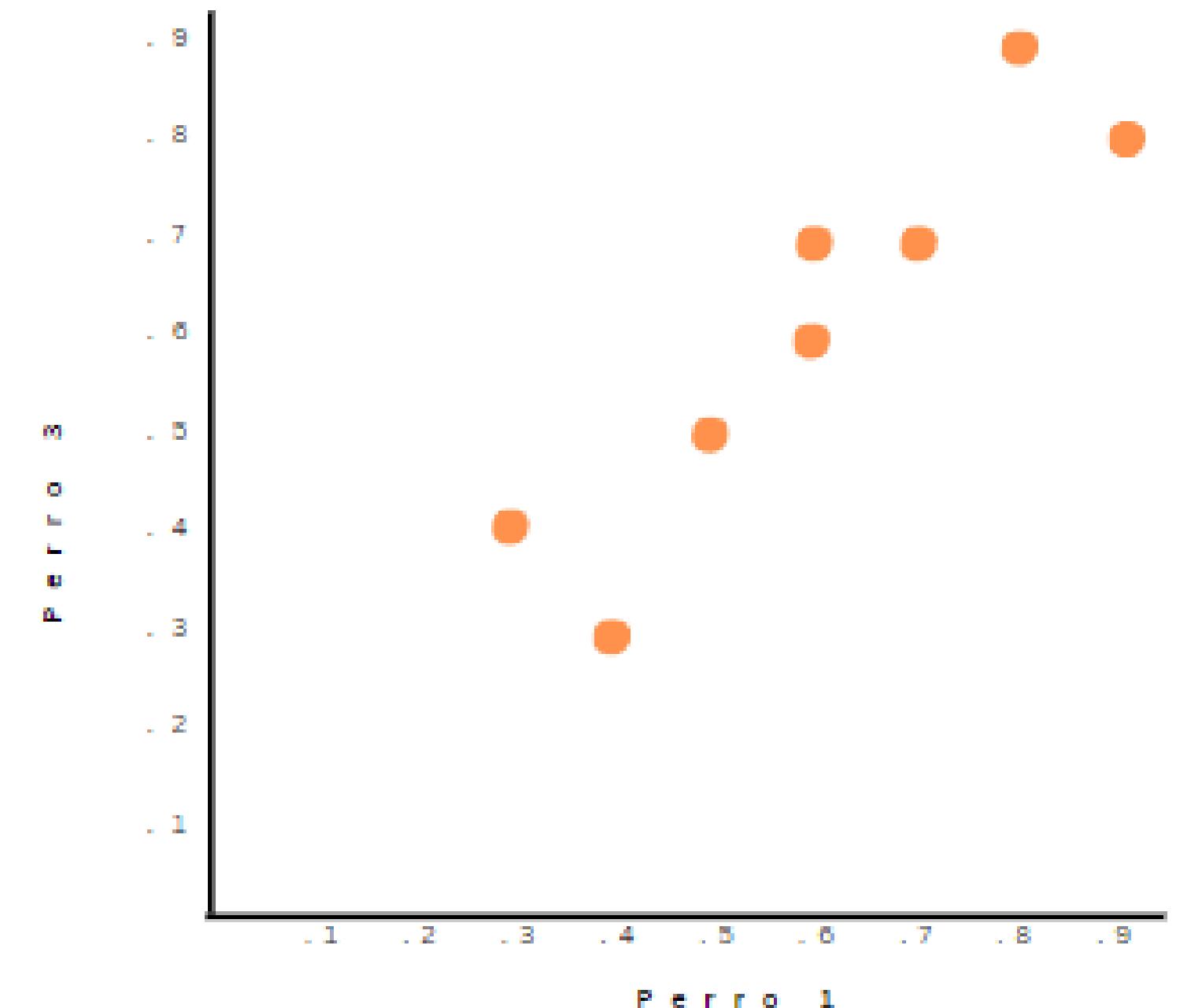
Diferentes razas



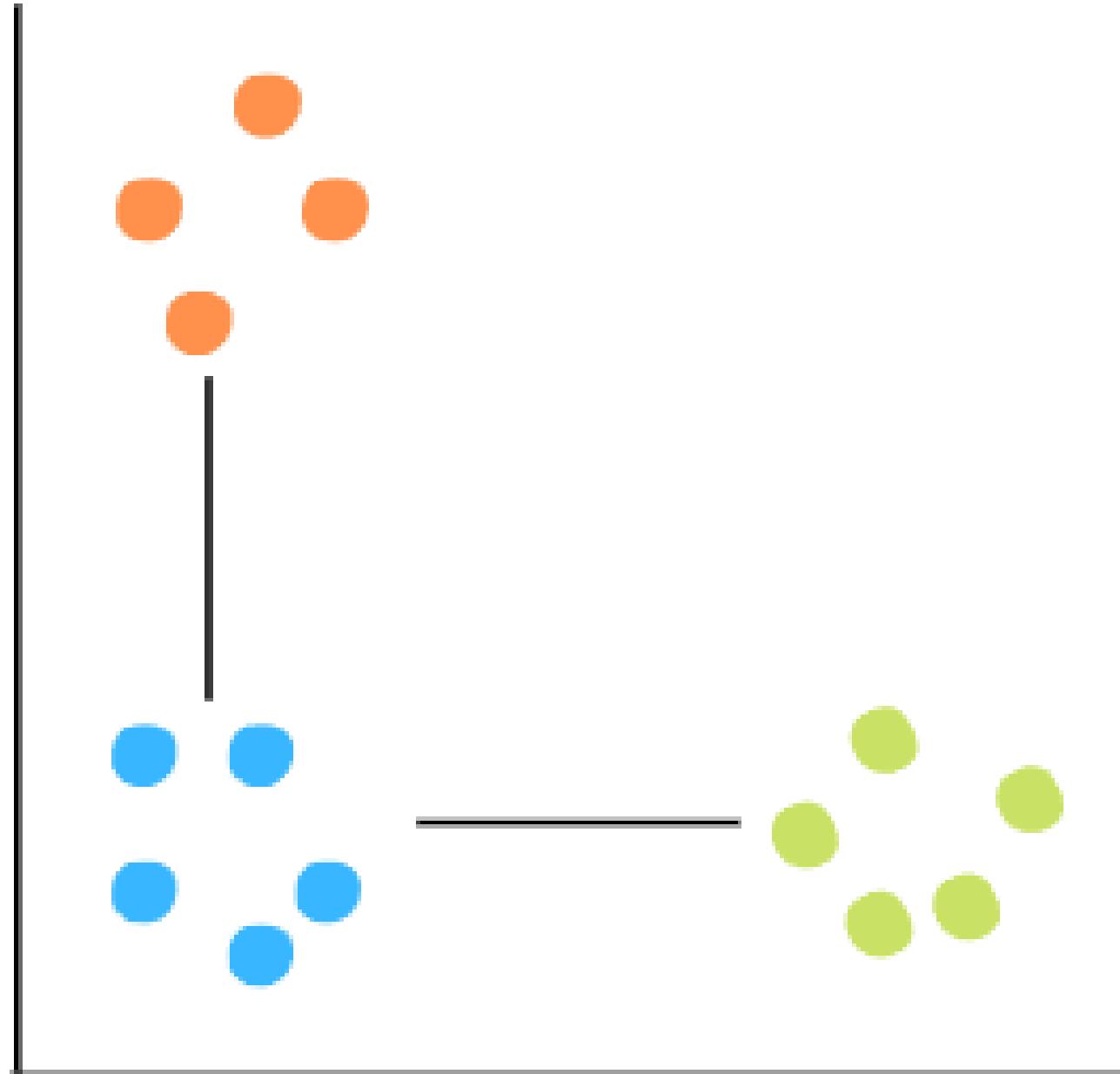
PCA (Principal Component Analysis)

Registro de raza	Perro1	Perro3
Gen1	0.8	0.9
Gen2	0.4	0.3
Gen3	0.7	0.7
Gen4	0.3	0.4
Gen5	0.6	0.7
Gen6	0.5	0.5
Gen7	0.6	0.6
Gen8	0.9	0.8

Misma raza



PCA (Principal Component Analysis)



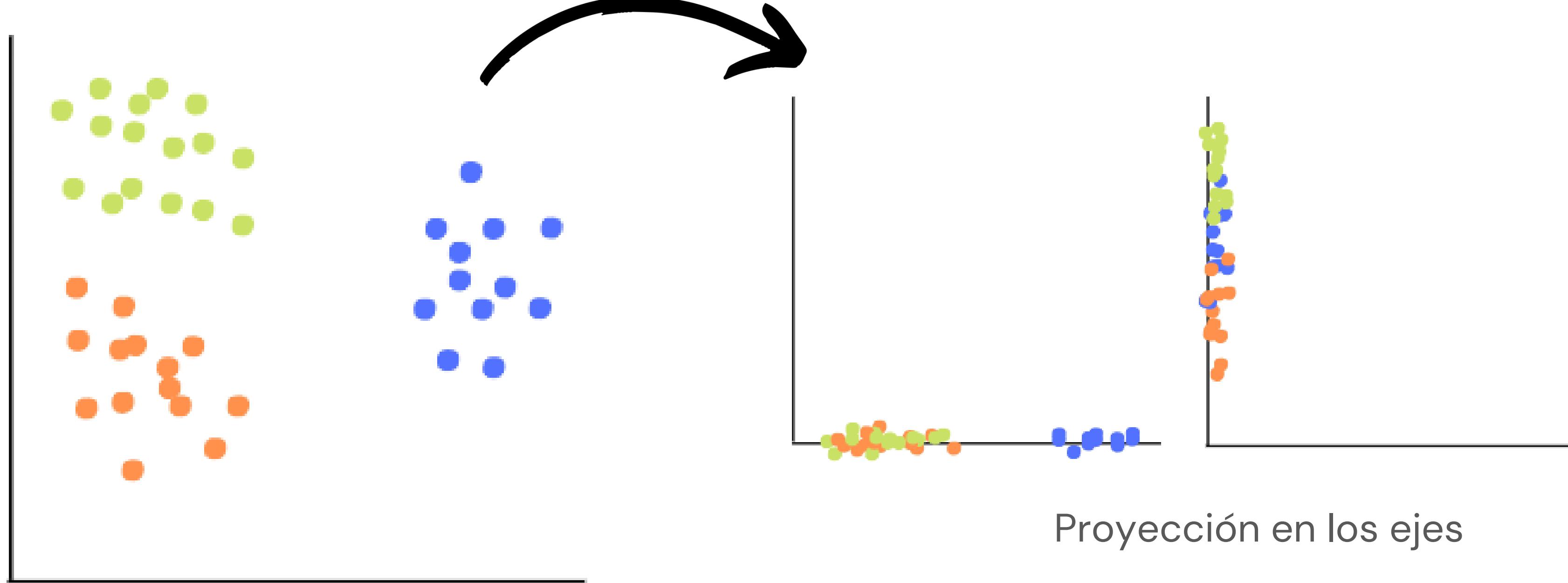
Las diferencias entre los componentes mirando desde cada eje son distintas.

Por el diagrama, podemos ver que el grupo azul está a la misma distancia del grupo naranja y del grupo verde. Pero, el grupo azul es “más diferente” del grupo verde, que haciendo la comparación con el grupo naranja

t-SNE (T-Distributed Stochastic Neighbor Embedding)

PCA puede encontrar relaciones lineales entre los datos y dar una vista global de la estructura de los datos.

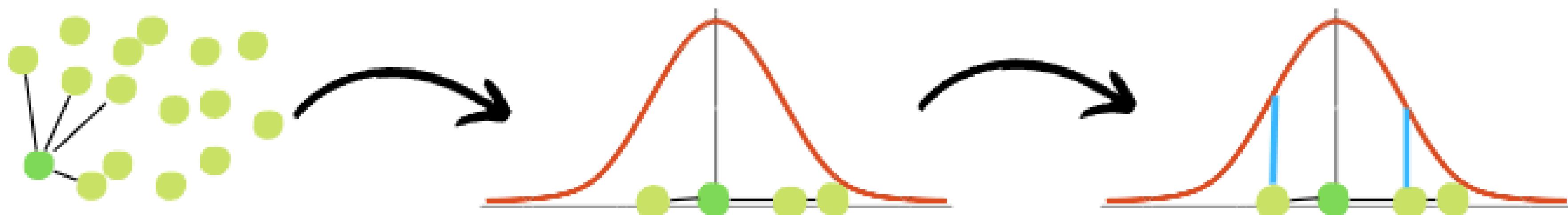
Pierdo información, ya no tengo grupos (clusters)



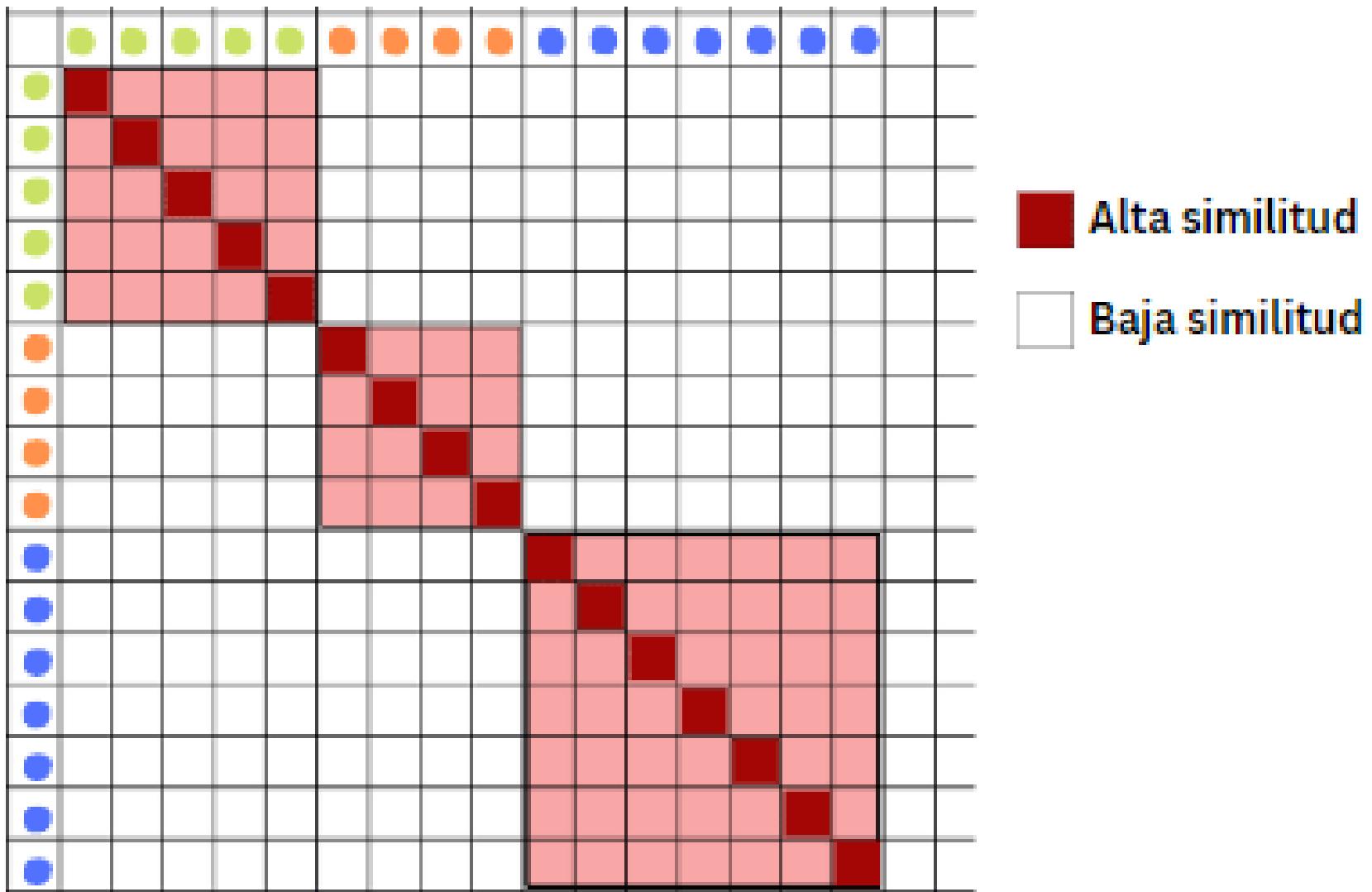
Parte uno

¿Qué se hace?

Se toma cada cluster. Una observación y se mide la distancia entre todas las observaciones del grupo y se reflejan en la curva normal. Luego se mide la distancia (en azul) de cada punto hasta que toca la curva.

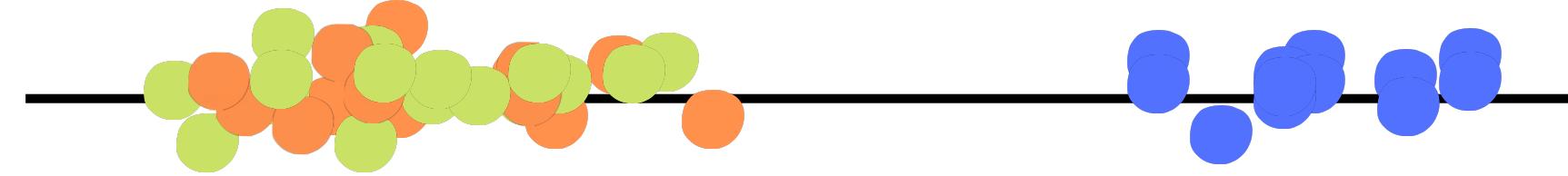


Parte uno



Obtenemos esta matriz con esas longitudes que nos habla sobre alta o baja similitud entre cada uno de los puntos

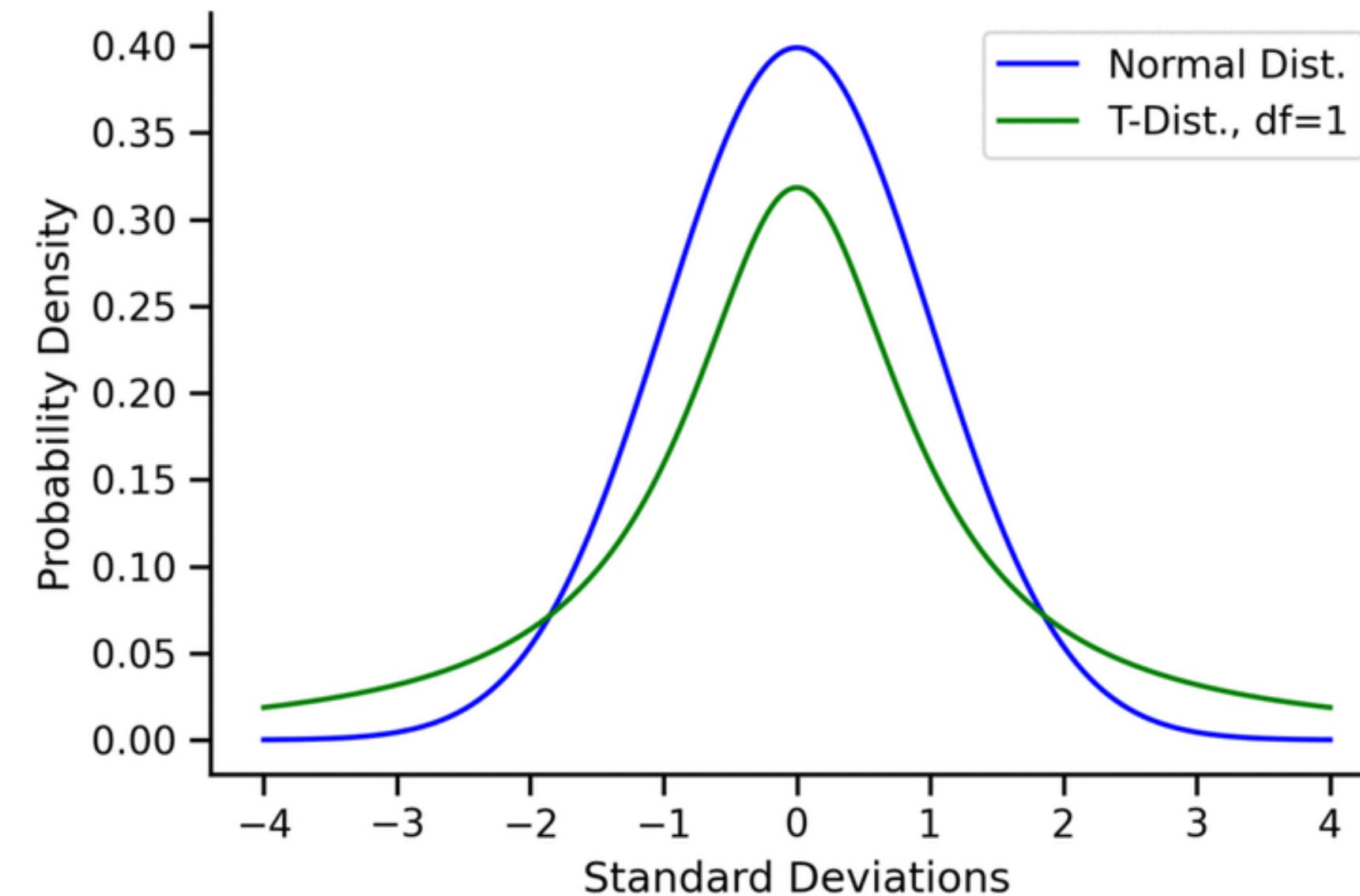
t-SNE (T-Distributed Stochastic Neighbor Embedding)



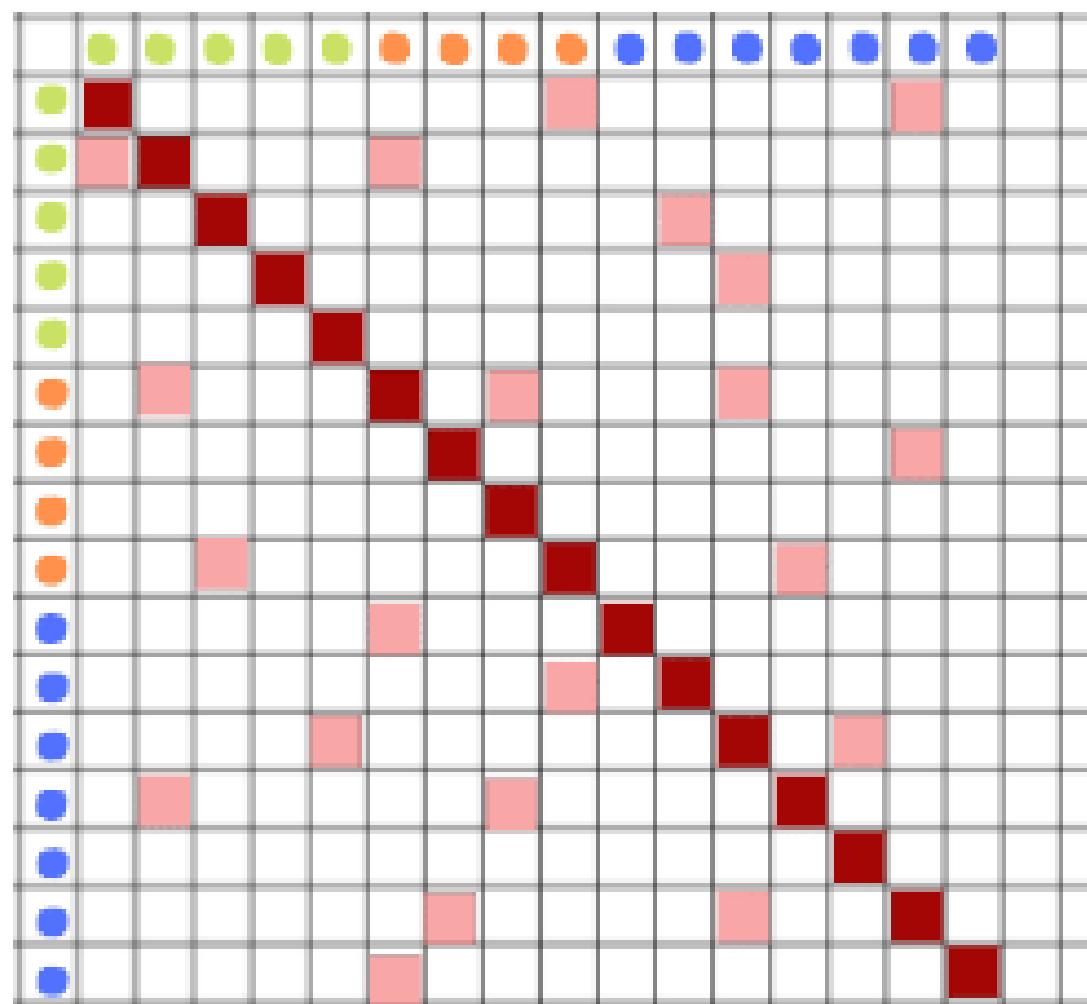
Parte dos

t-SNE (T-Distributed Stochastic Neighbor Embedding)

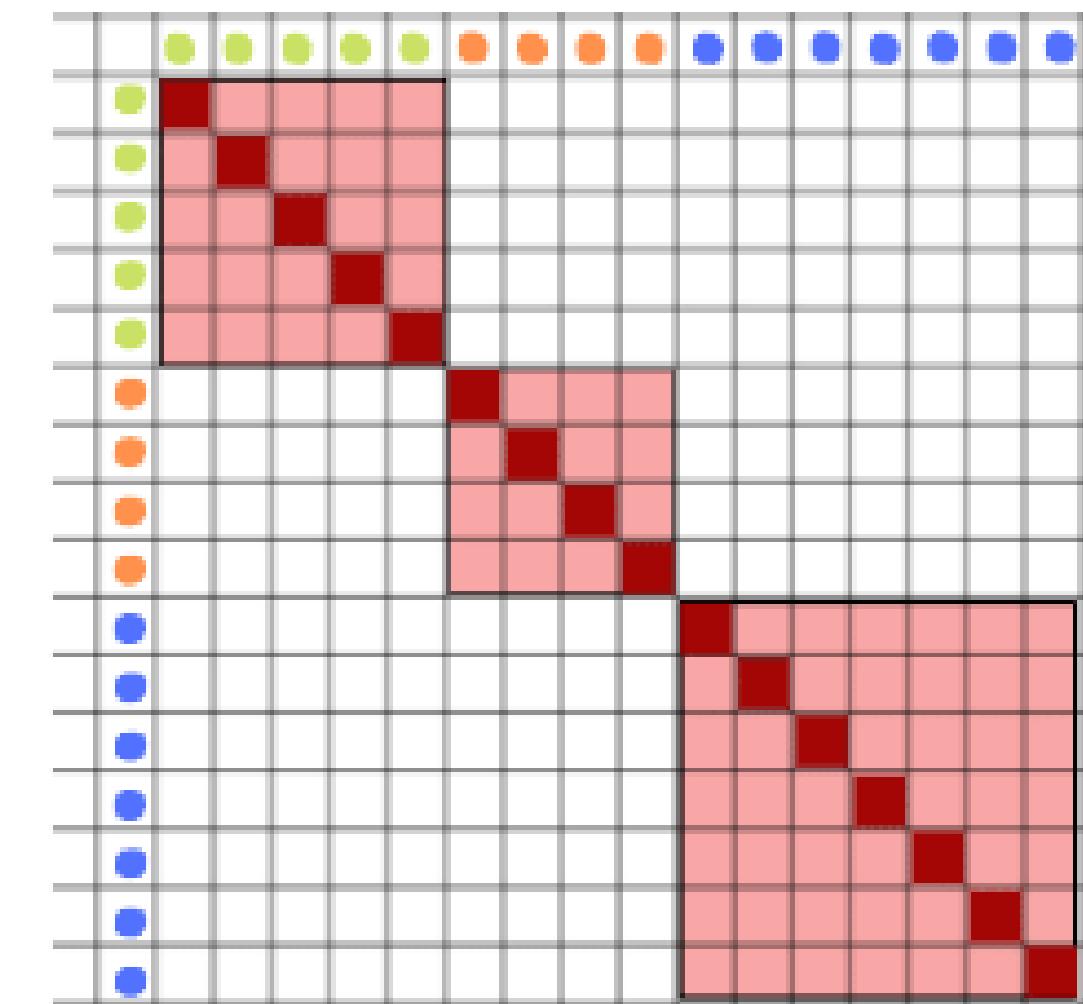
Parte dos



t-SNE (T-Distributed Stochastic Neighbor Embedding)

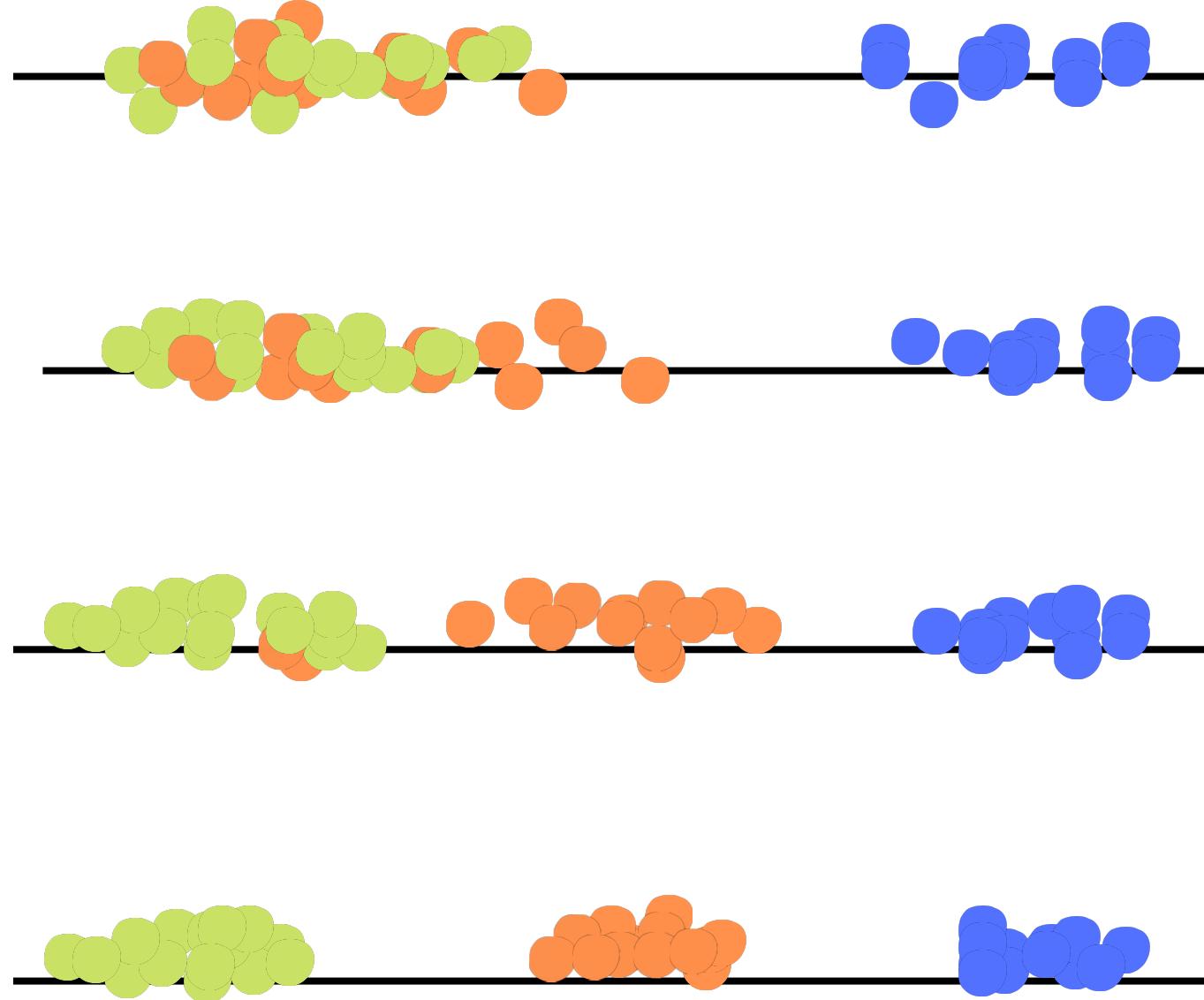


Usando t-student



Usando normal

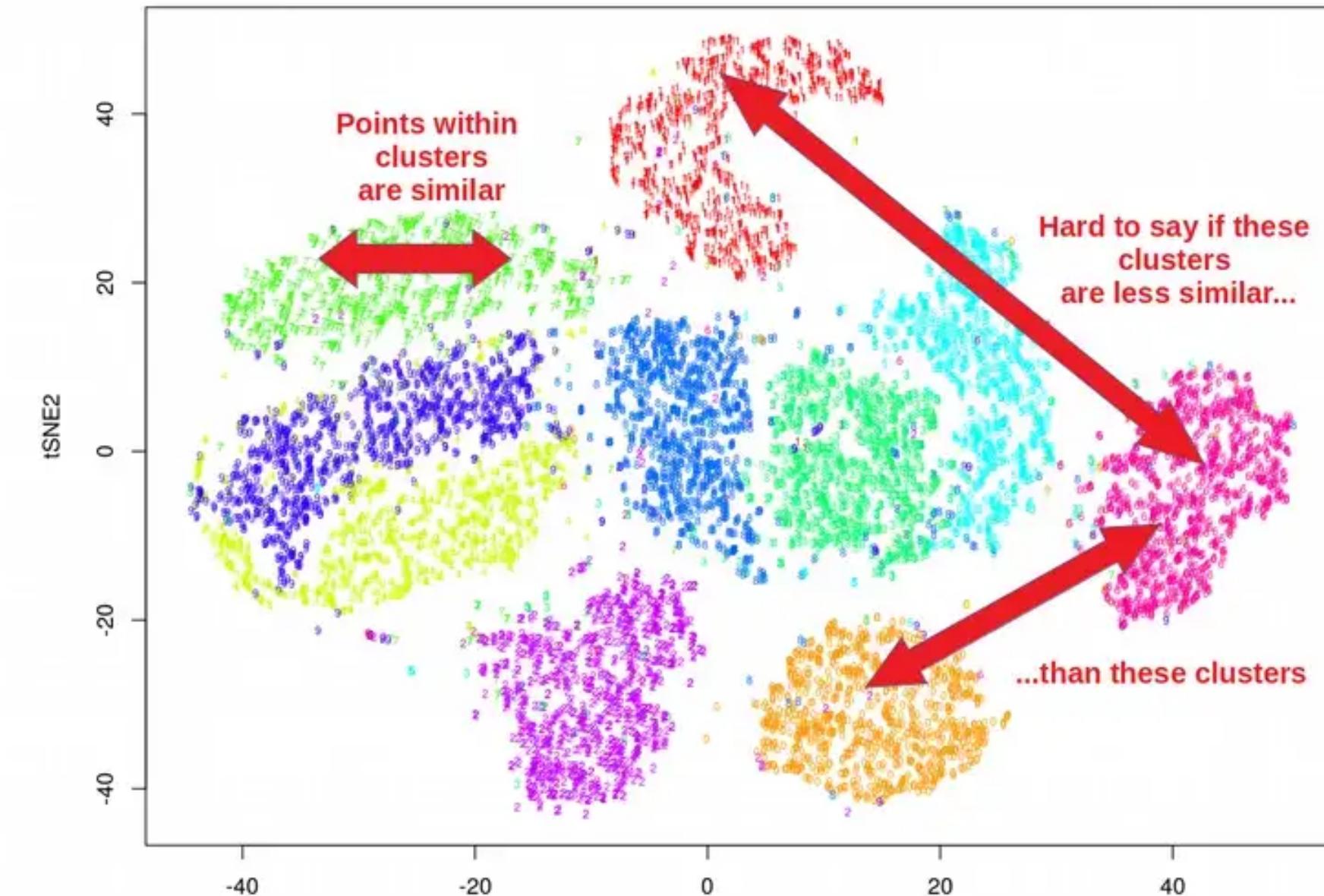
t-SNE (T-Distributed Stochastic Neighbor Embedding)



Parte dos

t-SNE (T-Distributed Stochastic Neighbor Embedding)

- No logra preservar la estructura general de los datos.
- No es un proceso rápido



Tomado de: <https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668>

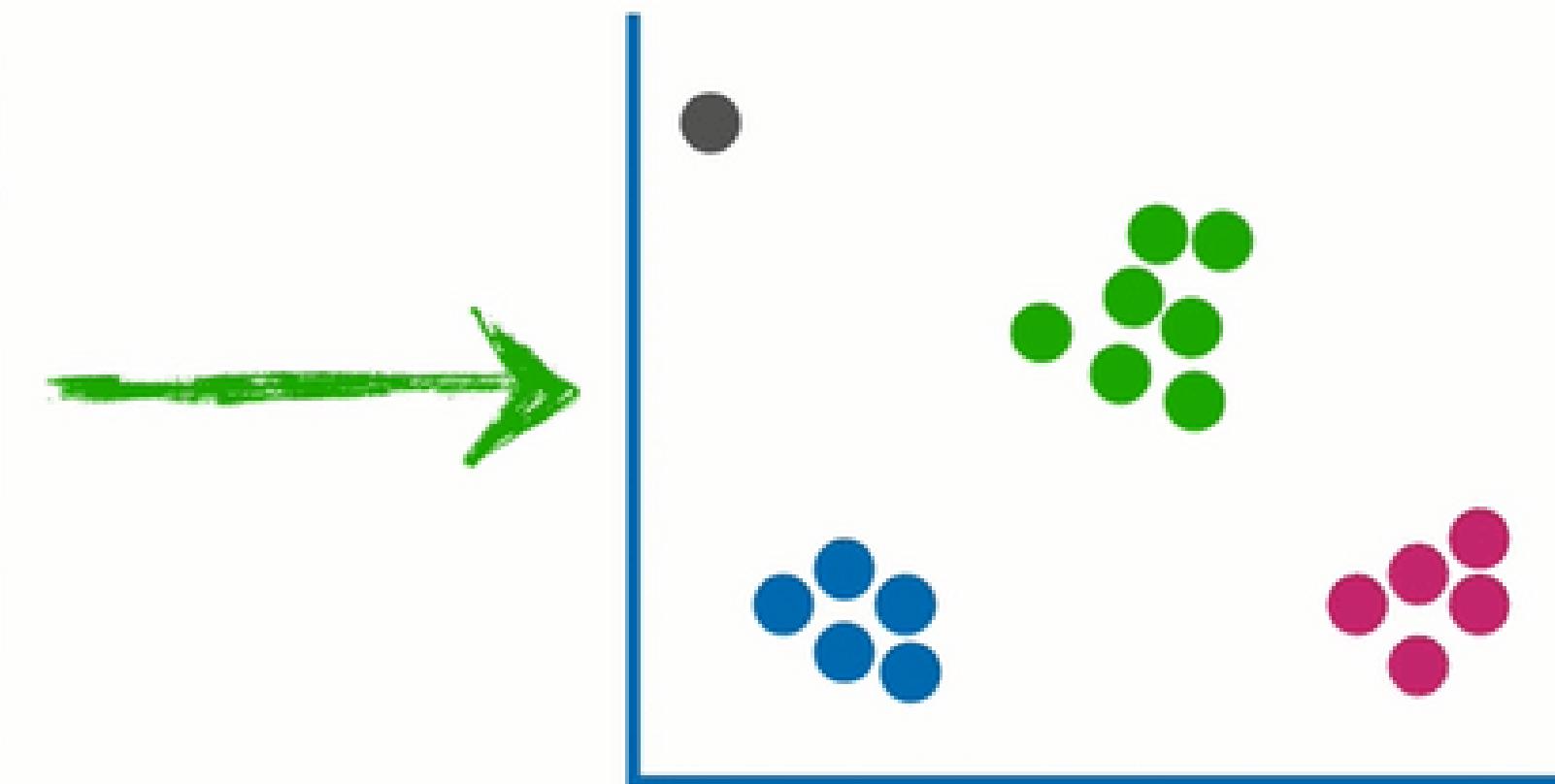
UMAP



- Es una técnica más rápida que t-SNE
- Identifica outliers.
- Preserva los grupos y su relación entre ellos y con otros.

UMAP

	Weight	Height	Age	...
Person 1	56	150	54	...
Person 2	62	170	21	...
Person 3	71	168	34	...
...

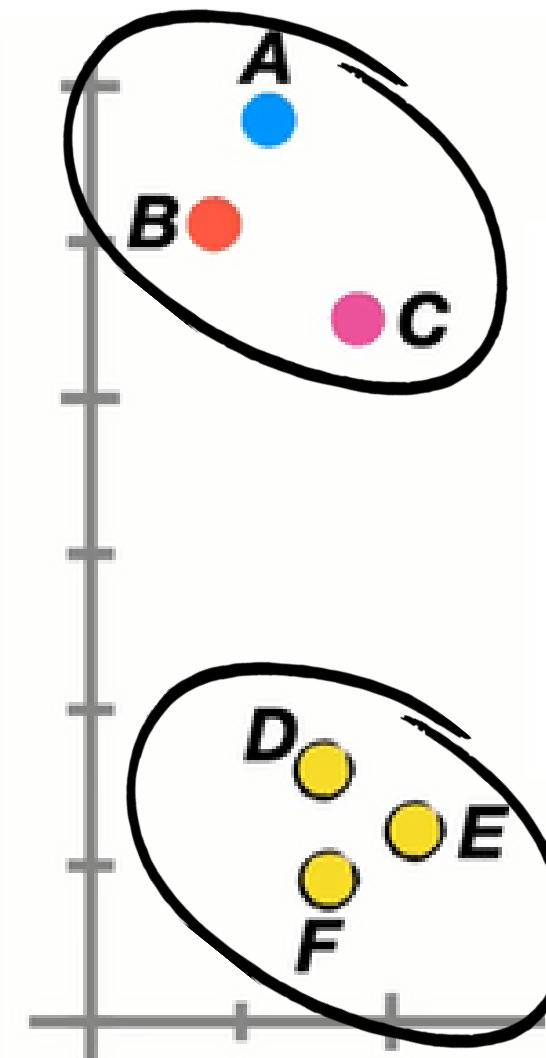


Identificar outliers y
similaridades

Tomado de:<https://www.youtube.com/watch?v=eNOwFzBA4Sc&t=2s>

UMAP

Datos de alta dimensión



Objetivo es lograr un gráfico de menor dimensionalidad que preserve los grupos (clusters) de los datos en alta dimensionalidad y su relación entre ellos

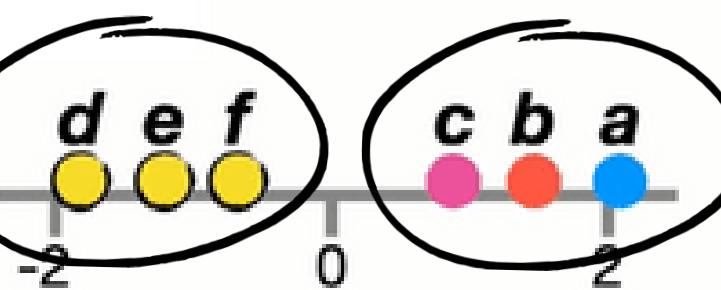


Gráfico de baja dimensión

Tomado de: <https://www.youtube.com/watch?v=eNOwFzBA4Sc&t=2s>



IDEAS PRINCIPALES

¿Cómo hacemos UMAP?
Fundamentación matemática

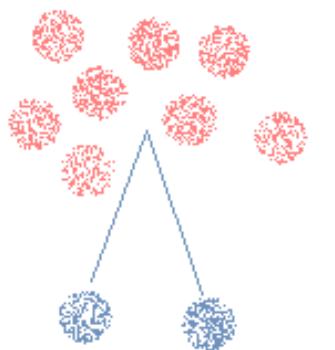
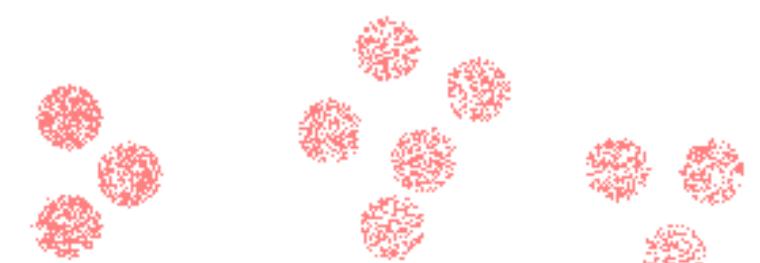


Minería de Datos
Pregrado en Estadística
Facultad de Ciencias - Sede Bogotá

¿Cómo hacemos UMAP?

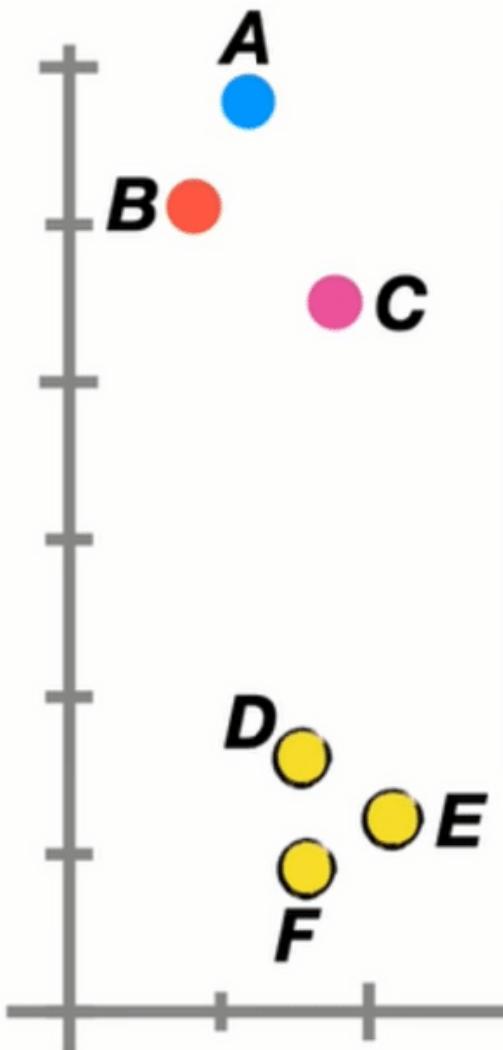
PROBLEMA: Representaciones topológicas difusas.

- I. Dar una **aproximación a la agrupación** (variedad) en el espacio de alta dimensionalidad del cual son originales los datos.
2. Hacer una **representación** de esta agrupación en un espacio de menor dimensión.

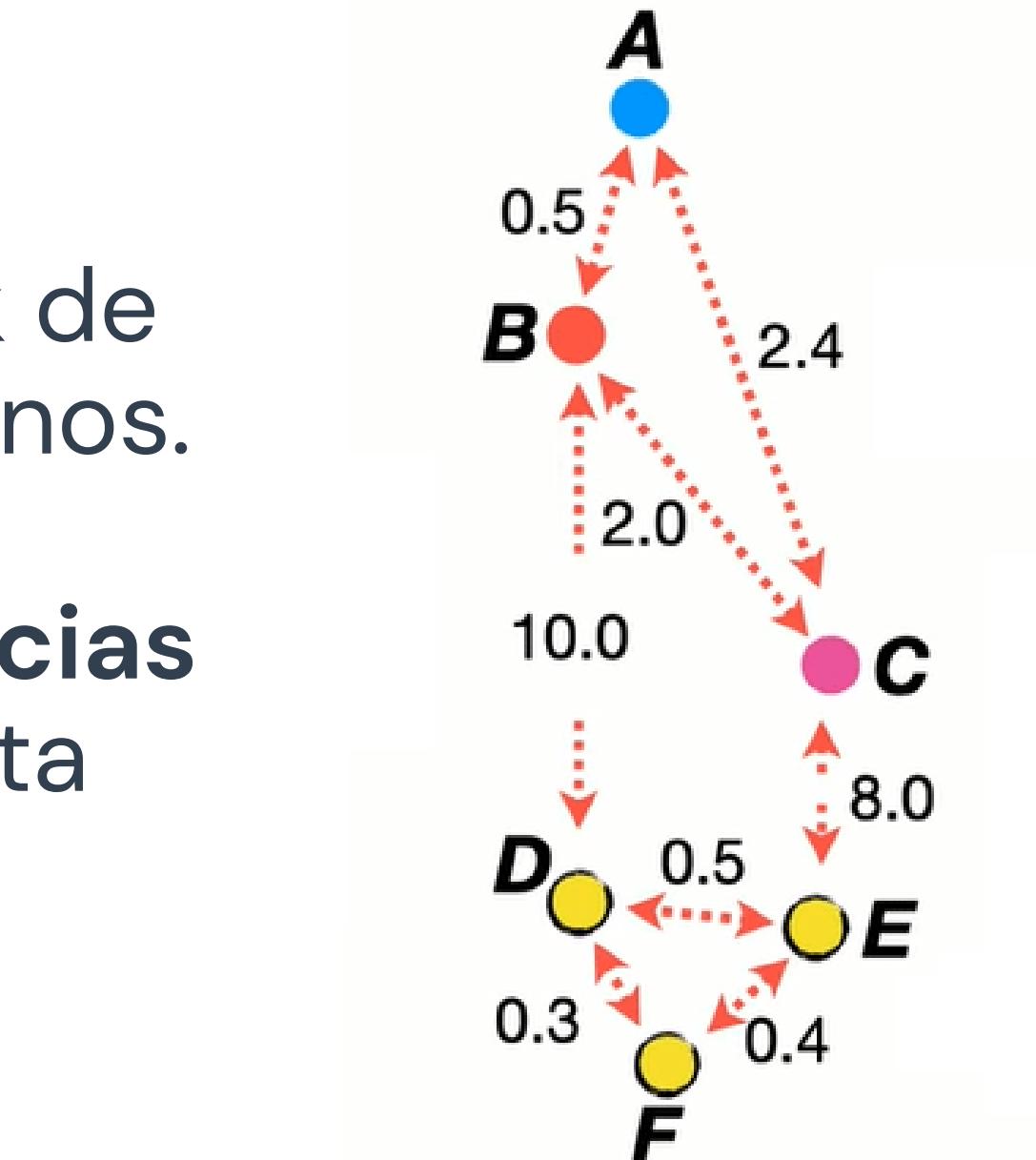


I. APROXIMACIÓN VARIEDAD

Datos de alta dimensión



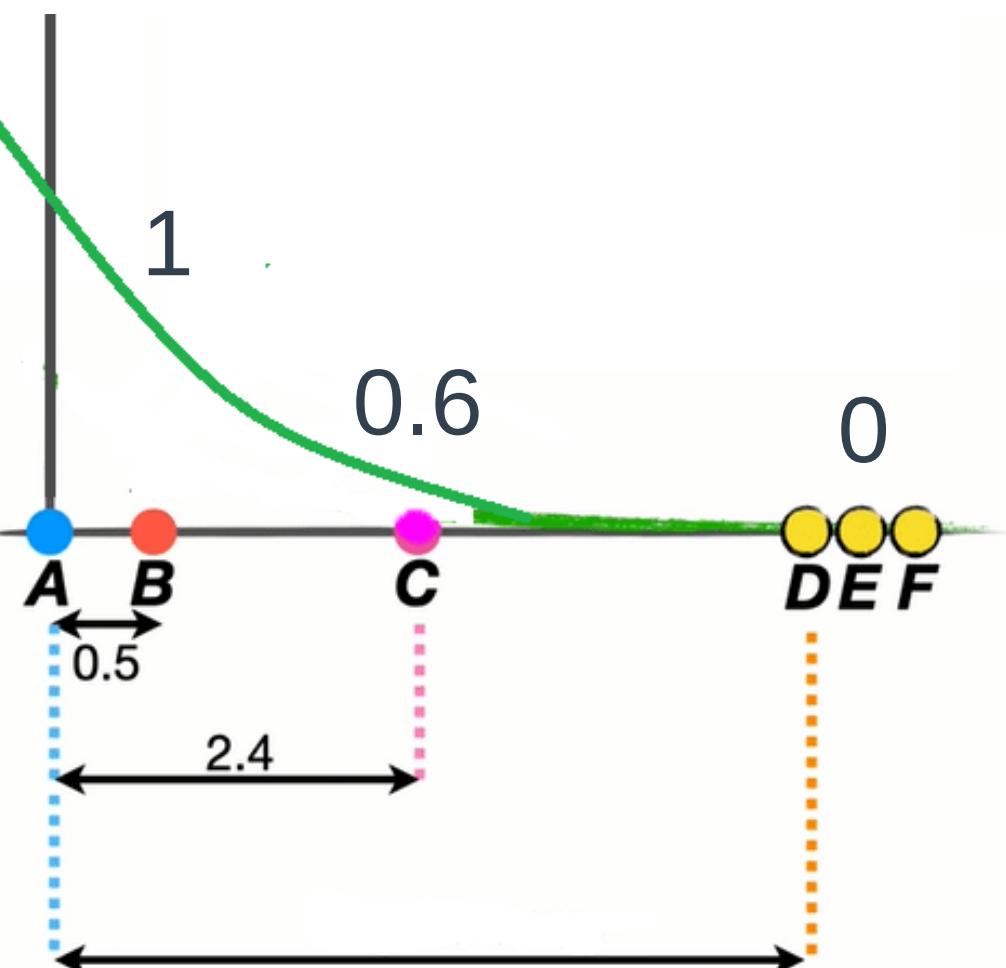
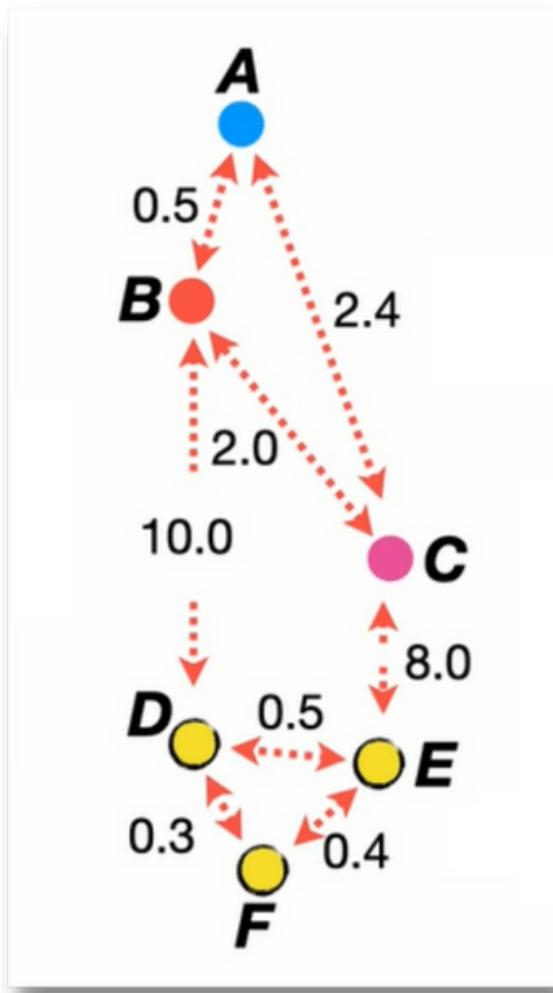
- Definir el número k de vecinos más cercanos.
- Calcular las **distancias** en el espacio de alta dimensión



I. APROXIMACIÓN VARIEDAD

k=3

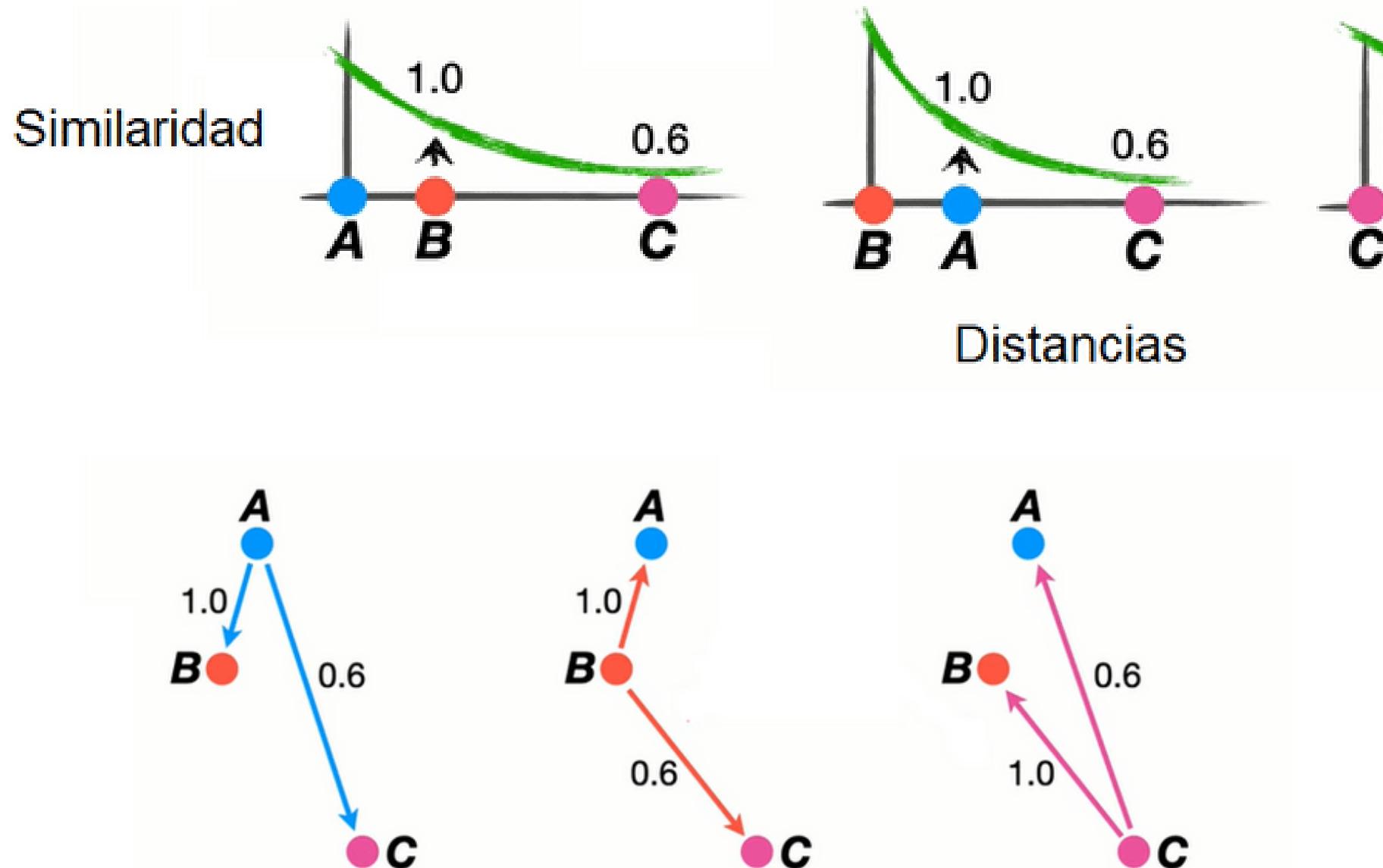
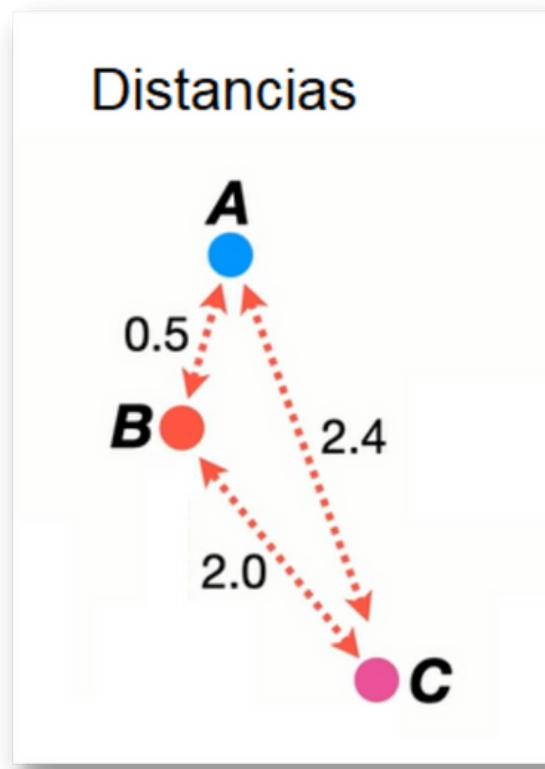
$$\log_2(\text{número de vecinos}) = \log_2(3) = 1.6$$



- Calcular los pesos de entrada.

SIMILARIDAD

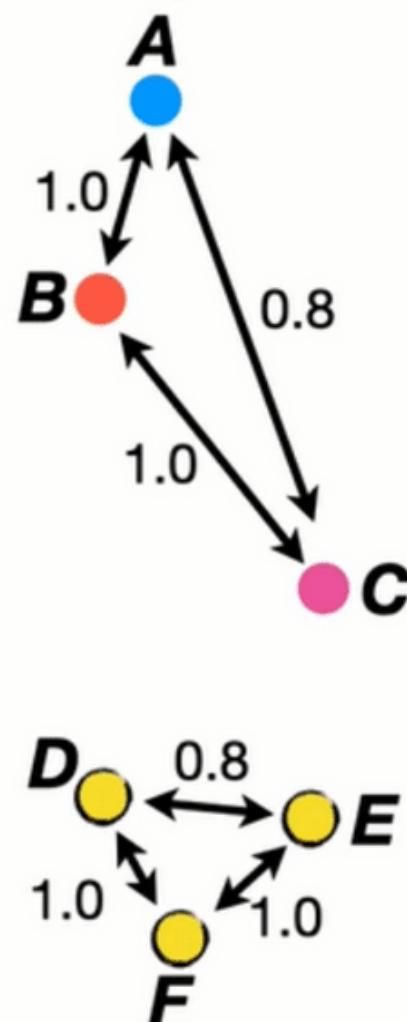
I. APROXIMACIÓN VARIEDAD



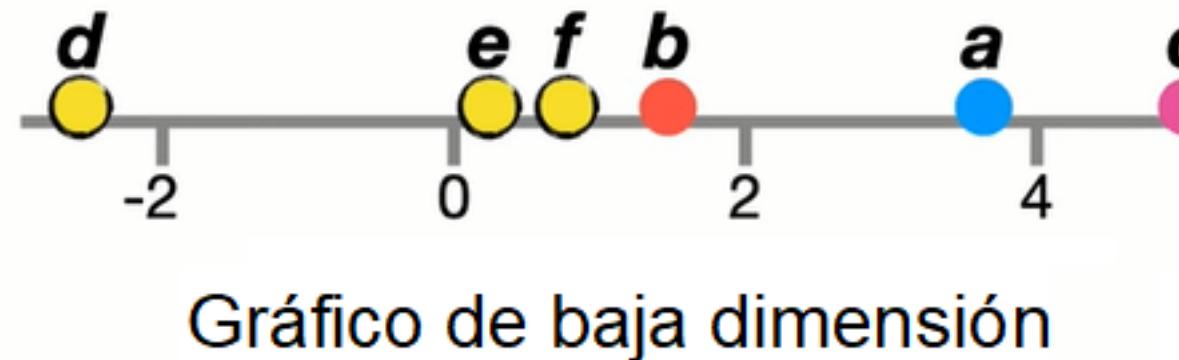
- Simetrizar los pesos de entrada
- Aproximar variedad

2. REPRESENTACIÓN

Pesos de entrada
simétricos



- Inicialización del gráfico de baja dimensión. Incrustación espectral



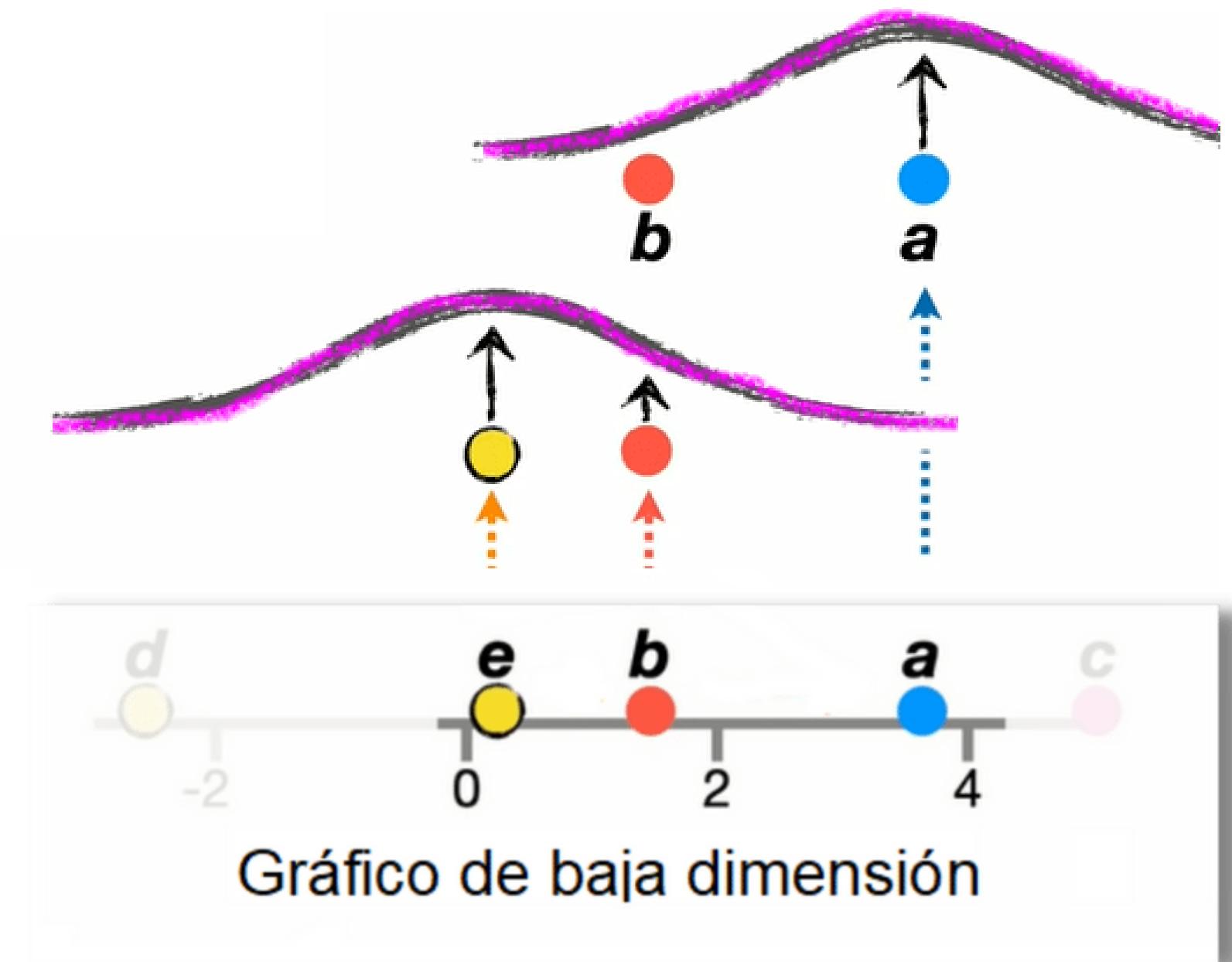
Optimizar:

- Escoger dos puntos en el mismo grupo para **acercar**.
- Escoger un punto externo para **alejarse**.

2. REPRESENTACIÓN

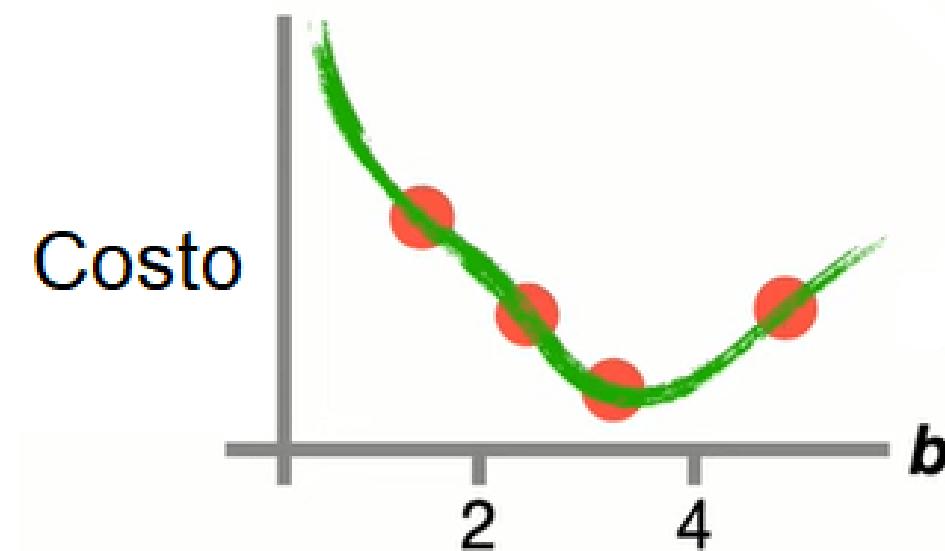
- Calcular los pesos de salida.

¿Qué tanto debo acercar o alejar **b**?



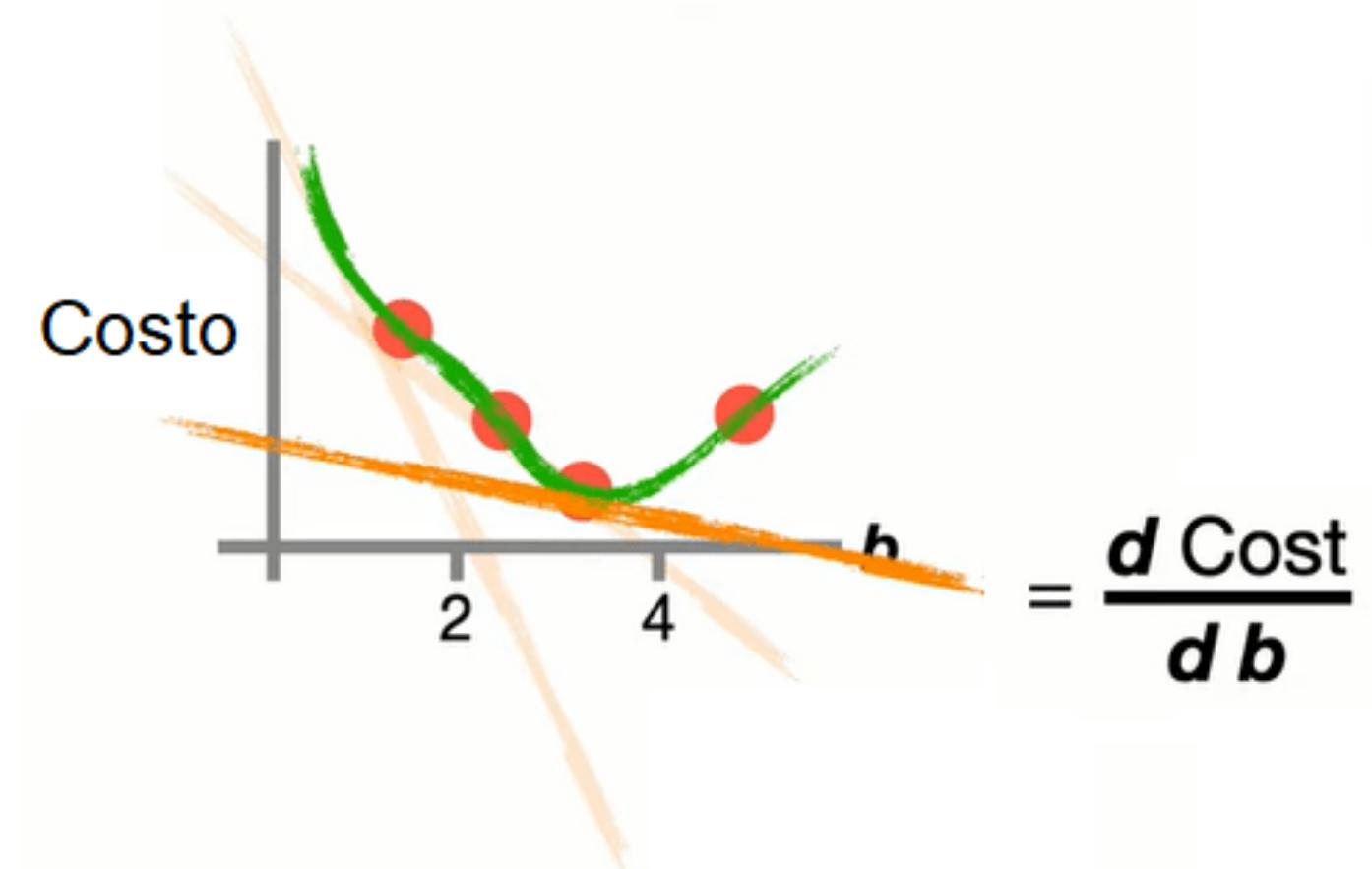
2. REPRESENTACIÓN

- Calcular la función de pérdida.



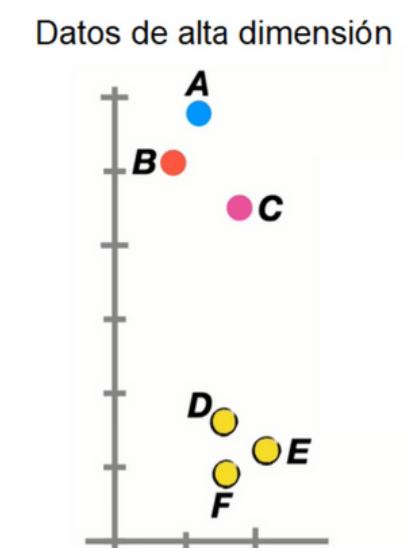
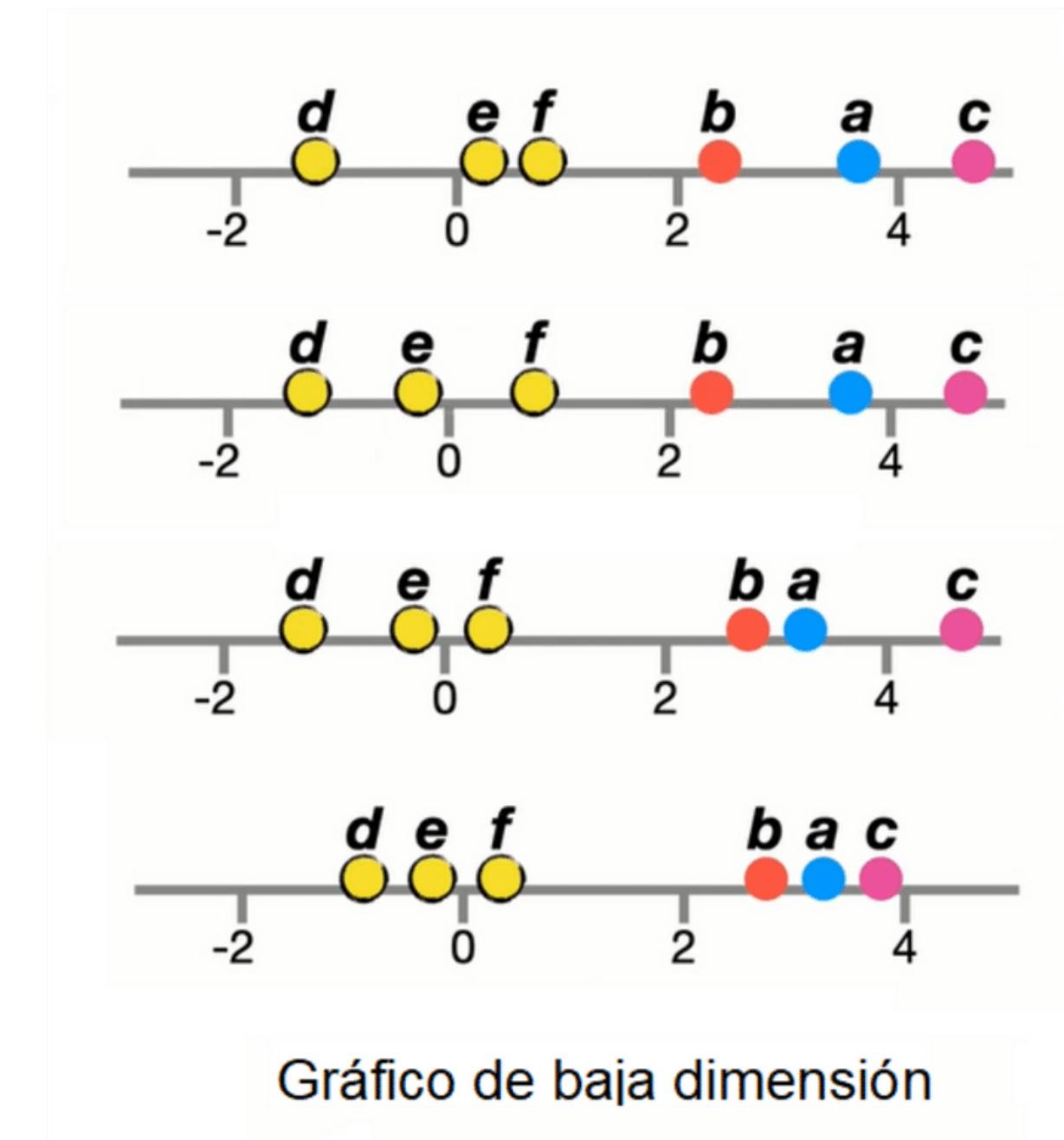
Equilibrio entre la cercanía
a a y la lejanía de e

- Descenso estocástico del gradiente



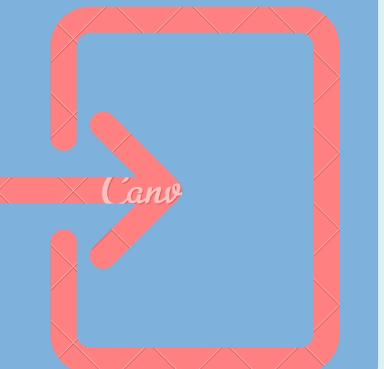
2. REPRESENTACIÓN

- Muevo uno o un pequeño grupo de puntos en cada fase.
- Aunque tienen el mismo "inicio", la **aleatoriedad del SGD** hace variar la representación final.

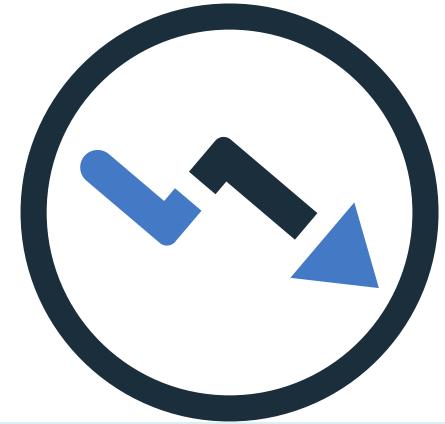
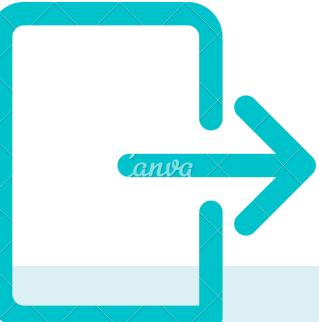


I. APROXIMACIÓN VARIEDAD

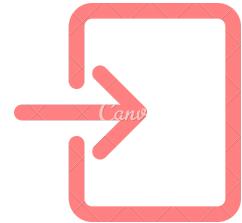
- Definir el número k de vecinos más cercanos.
- Calcular las **distancias** en el espacio de alta dimensión.
- Calcular los **pesos de entrada**.
(Similitud)
- Simetrizar los pesos de entrada



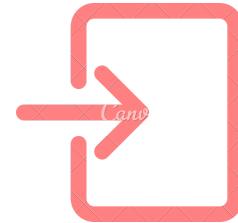
- Iniciar la representación.
- Calcular los **pesos de salida**.
- Calcular la **función de pérdida**.



2. REPRESENTACIÓN



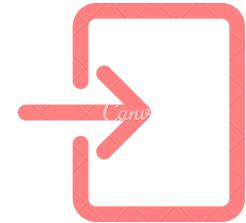
PESOS DE ENTRADA



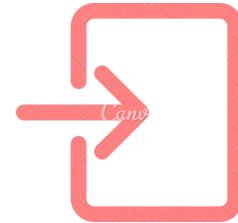
Dadas N observaciones de un conjunto de datos de alta dimensión $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$

$$v_{j|i} = \exp[-(r_{ij} - \rho_i) / \sigma_i]$$

- r_{ij} es la distancia entre \mathbf{x}_i y \mathbf{x}_j .
- ρ_i es la distancia la observación \mathbf{x}_i a su vecino más cercano.
- σ_i es una constante determinada para cumplir $\sum_j v_{j|i} = \log_2 k$, donde k es el número de vecinos más cercanos.
- $v_{j|i} \in (0, 1]$. Note que $v_{j|i} = 1$ cuando \mathbf{x}_j es el vecino más cercano a \mathbf{x}_i y disminuye a medida que la observación se encuentre más lejos de \mathbf{x}_i .



PESOS DE ENTRADA



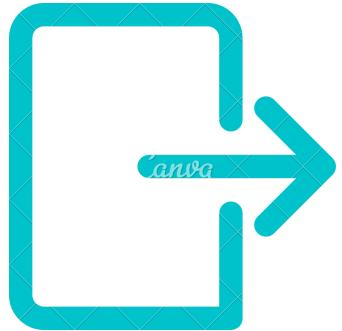
$$v_{j|i} = \exp[-(r_{ij} - \rho_i) / \sigma_i]$$

SIMETRIZACIÓN

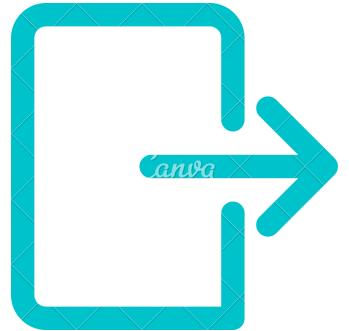
La notación $v_{j|i}$ en lugar de v_{ij} se usa para representar la relación asimétrica, es decir, $v_{j|i} \neq v_{i|j}$. Sin embargo, es conveniente simetrizar esta similaridad mediante

$$v_{ij} = (v_{j|i} + v_{i|j}) - v_{j|i}v_{i|j}$$

la cual lleva a cabo efectivamente una *unión de conjuntos difusos*.



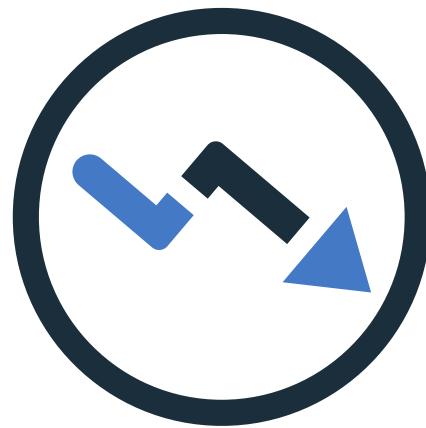
PESOS DE SALIDA



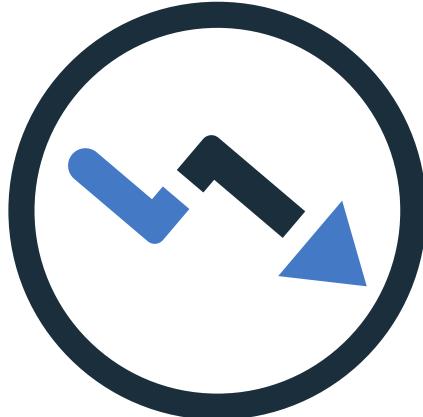
INICIO coordenadas incrustadas $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}$

$$w_{ij} = 1 / \left(1 + \alpha d_{ij}^{2\beta} \right)$$

- d_{ij} es la distancia euclídea entre \mathbf{y}_i e \mathbf{y}_j .
- α, β son constantes que controlan la rigidez de la función de agrupamiento.
- $w_{ij} \in (0, 1]$. Note que $w_{ij} = 1$ cuando \mathbf{y}_j tiene las mismas coordenadas de \mathbf{y}_i y disminuye a medida que la observación se encuentre más lejos de \mathbf{y}_i .



FUNCIÓN DE PÉRDIDA



Entropía cruzada de dos conjuntos borrosos, la cual se minimiza por componentes para encontrar la posición de cada observación.

$$C_{UMAP} = \sum_{i,j} \left[v_{ij} \log\left(\frac{v_{ij}}{w_{ij}}\right) + (1 - v_{ij}) \log\left(\frac{1 - v_{ij}}{1 - w_{ij}}\right) \right]$$



APLICACION

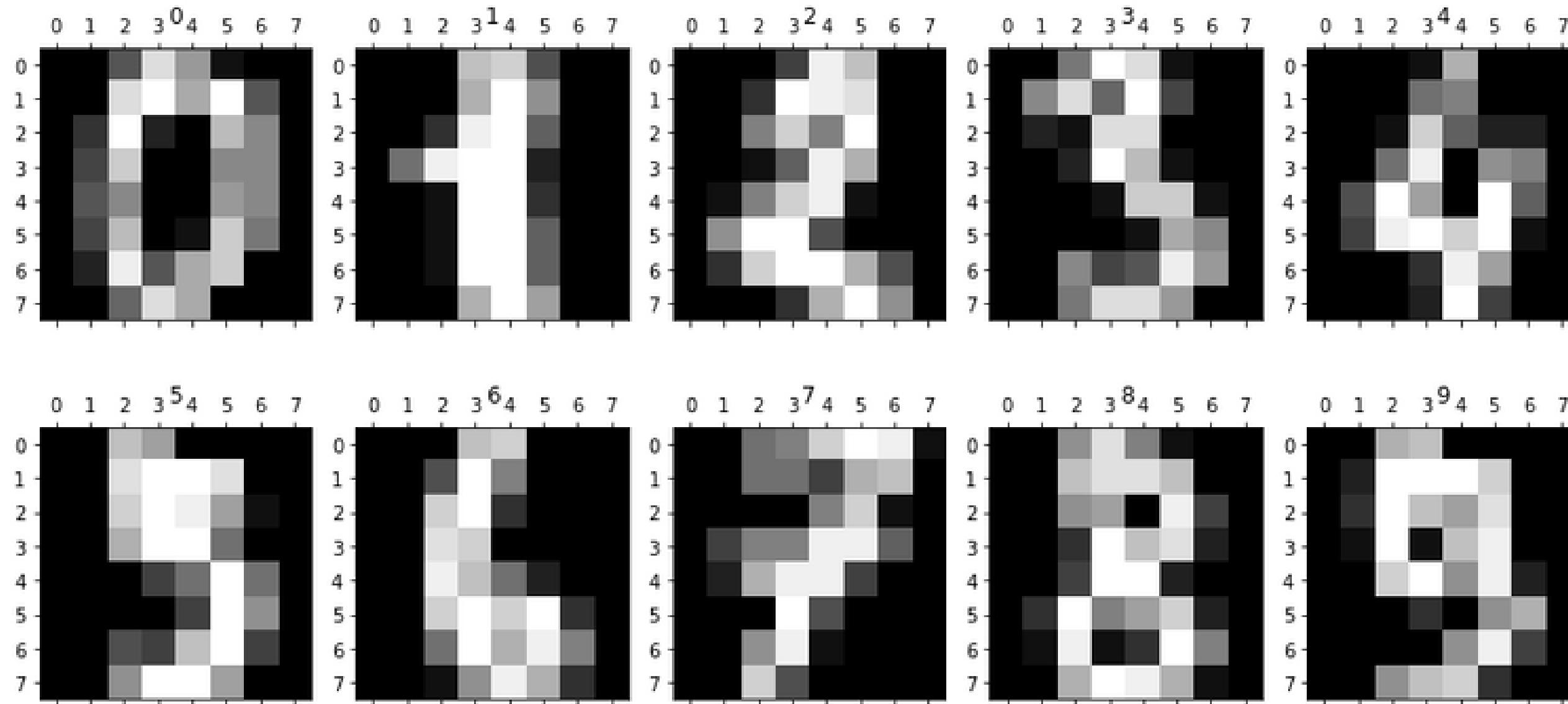


Minería de Datos
Pregrado en Estadística
Facultad de Ciencias - Sede Bogotá



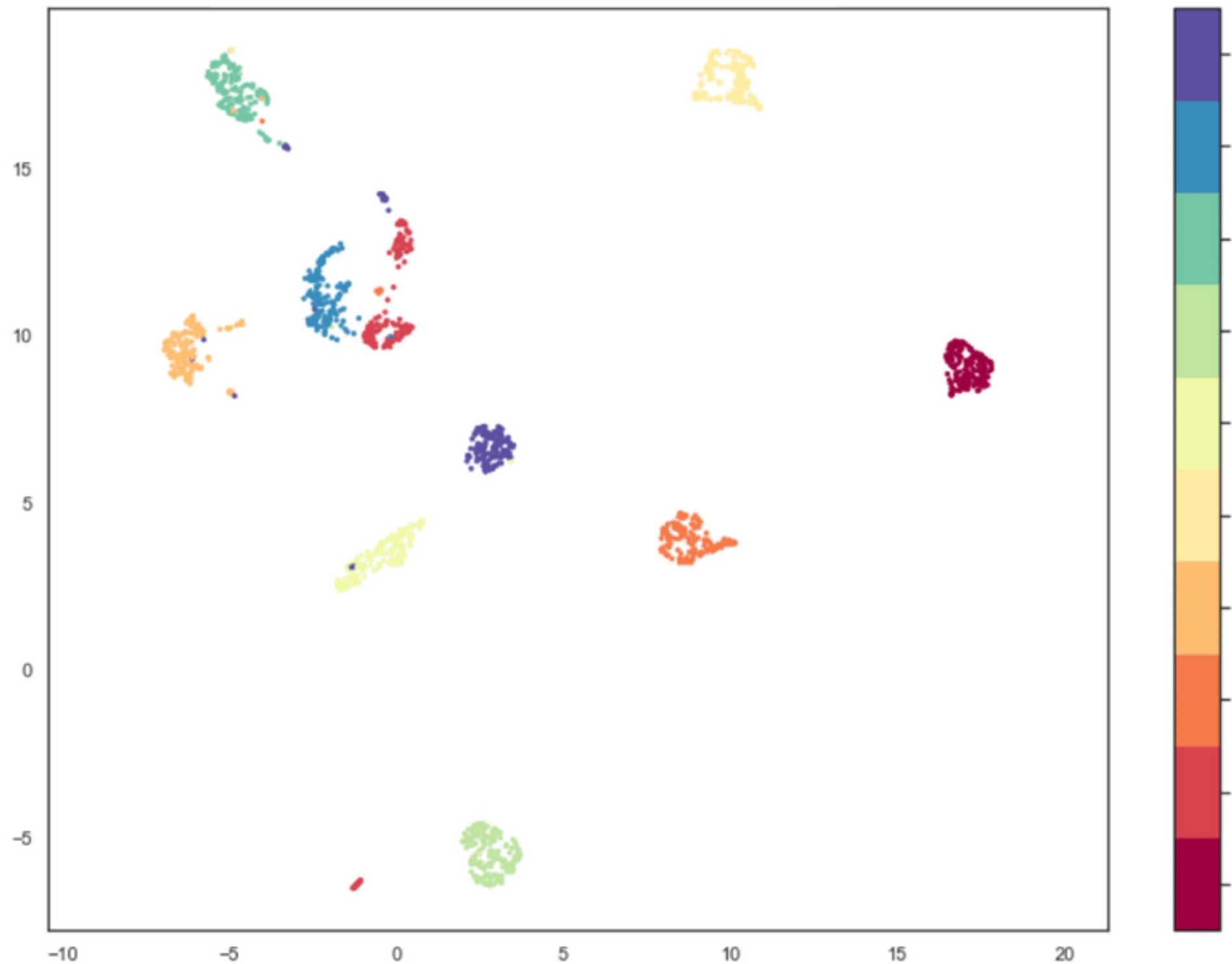
UNIVERSIDAD
NACIONAL
DE COLOMBIA

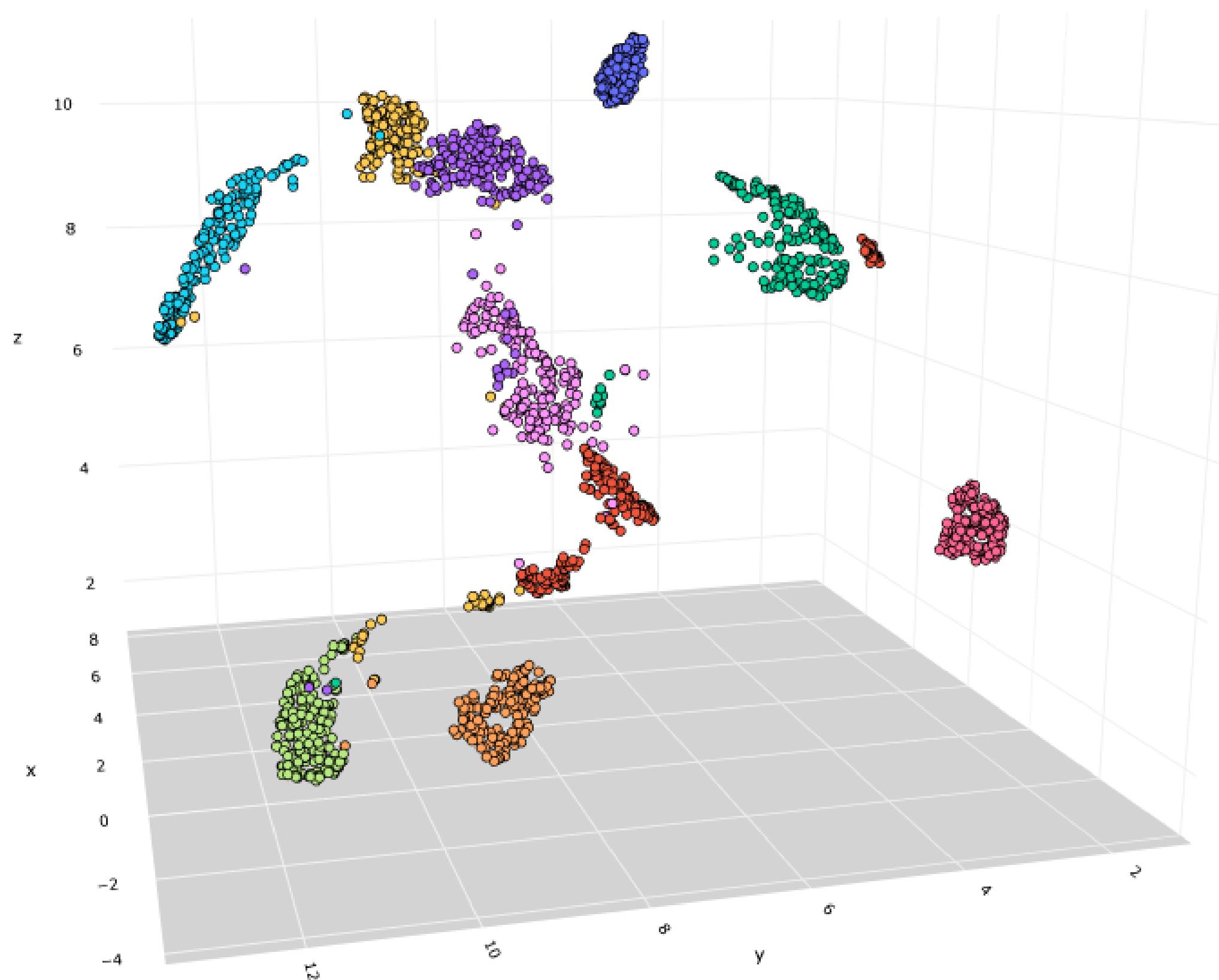
mnist



(1797, 8, 8)

UMAP para la base de datos de díaitos

**2 DIMENSIONES**



3 DIMENSIONES

MUCHAS GRACIAS



Nicolle Stefania Quintero Motta
nquinterom@unal.edu.co

Paula Camila Garcia Nieto
pagarcian@unal.edu.co

Noviembre 2022

*Minería de Datos
Pregrado en Estadística
Facultad de Ciencias - Sede Bogotá*



REFERENCIAS

Josh Starmer. [StatQuest] (2017, 18 de septiembre). t-SNE, Clearly Explained [video]. Youtube. <https://www.youtube.com/watch?v=NEaUSP4YerM>

Josh Starmer. [StatQuest] (2022, 7 de marzo). UMAP Dimension Reduction, Main Ideas!!! [video]. Youtube. <https://www.youtube.com/watch?v=eN0wFzBA4Sc&t=200s>

Josh Starmer. [StatQuest] (2017, 4 de diciembre). Principal ideas of PCA in 5 minutes [video]. Youtube. https://www.youtube.com/watch?v=HMOI_IkzW08&t=130s

McInnes, L, Healy, J, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018. <https://umap-learn.readthedocs.io/en/latest/api.html>

Oskolkov Nikolay. Towards Data Science (2019, 3 de octubre) ¿Cómo funciona UMAP y por qué es mejor que t-SNE? <https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668>

Pedregosa,F et al. Journal of Machine Learning Research (2011) <https://scikit-learn.org/stable/about.html#citing-scikit-learn>

JavaTpoint. Introduction to Dimensionality Reduction Technique <https://www.javatpoint.com/dimensionality-reduction-technique>