

# **Herramientas matemáticas y computacionales para la solución de problemas en Ciencia de Datos**

Odin Eufracio

Laboral



Reuters

## El Reino Unido compensará con robótica la marcha de trabajadores por el 'brexit'

5D  
EFE

- KPMG apunta a un aumento de la automatización, los robots y la inteligencia artificial
- Un 35% de los europeos que trabajan en el Reino Unido se plantean irse

[https://cincodias.elpais.com/cincodias/2017/09/02/mercados/1504345291\\_069578.html](https://cincodias.elpais.com/cincodias/2017/09/02/mercados/1504345291_069578.html)

# Cambridge Analytica, el big data y su influencia en las elecciones



Aníbal García Fernández

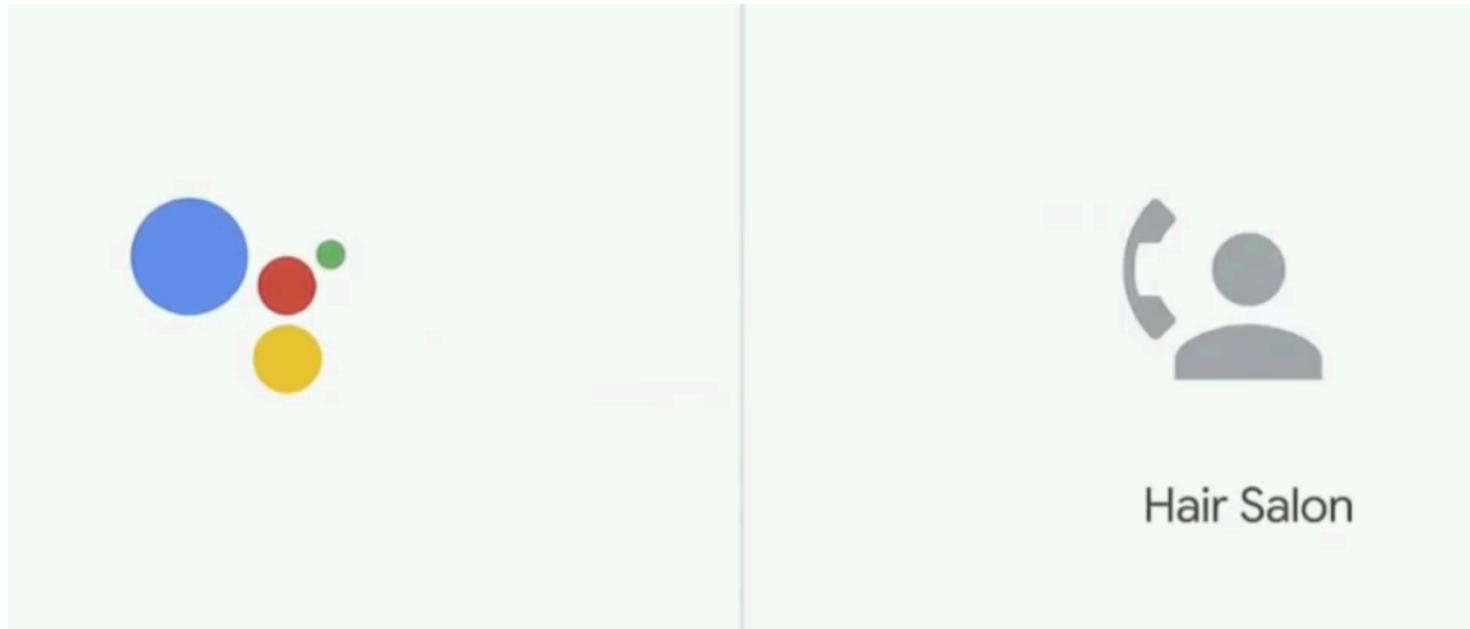


## Lo más leído

- *La economía boliviana en 2019*
- *Latinoamérica, más progresista*
- *Radiografía de la situación económica boliviana*
- *Argentina: un proyecto popular, democrático y latinoamericano*
- *20 años de dolarización en Ecuador: ¿es sostenible?*
- *Evo gana y Bolivia también*
- *Cambridge Analytica, el big data y su influencia en las elecciones*
- *Libre cambio Vs Proteccionismo. El doble estándar de los países desarrollados*
- *Bolivia y la revolución cultural*

<https://www.celag.org/cambridge-analytica-el-big-data-y-su-influencia-en-las-elecciones/>

# Google Duplex: A.I. Assistant Calls Local Businesses To Make Appointments



<https://www.youtube.com/watch?v=D5VN56jQMWM>

Presentar un ejemplo que nos permita ver la aplicación de las herramientas matemáticas y computacionales en la solución de un problema *clásico* en ciencia de datos

# Premio Netflix

# Premio Netflix

The screenshot shows the Netflix Prize website with a prominent yellow banner at the top featuring the text "Netflix Prize" and a large red "COMPLETED" stamp. Below the banner, there's a navigation bar with links for "Home", "Rules", "Leaderboard", and "Update". A main content area displays a "Movies For You" section with movie recommendations and a "You really liked it..." sidebar. A large white callout box in the center-right contains the word "Congratulations!" in blue. Below it, text explains the purpose of the prize and mentions the awarding of the \$1M Grand Prize to BellKor's Pragmatic Chaos team. It also encourages users to explore the Leaderboard and Forum. At the bottom, there are links for "FAQ", "Forum", and "Netflix Home", along with a copyright notice for 1997-2009.

**NETFLIX**

**Netflix Prize**

**COMPLETED**

Home | Rules | Leaderboard | Update

Movies For You

Randy, the following movies were chosen based on your interest in:

[Howling the Columbine](#)  
[Camryn: Season 1](#)  
[Camryn: Season 2](#)

**You really liked it...**

Now own it for just \$5.99

[Shop more titles](#)

**Congratulations!**

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the \$1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about [their algorithm](#), checkout team scores on the [Leaderboard](#), and join the discussions on the [Forum](#).

We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

FAQ | Forum | Netflix Home

© 1997-2009 Netflix, Inc. All rights reserved.

# Premio Netflix



	5				4
				3	
	1				2
	2			5	
	4		5		

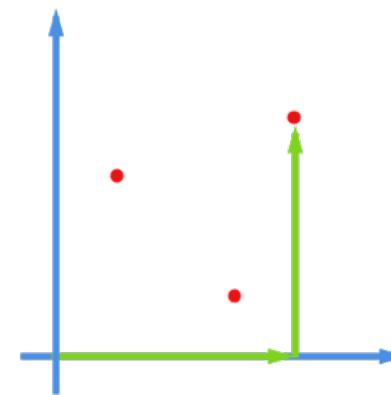
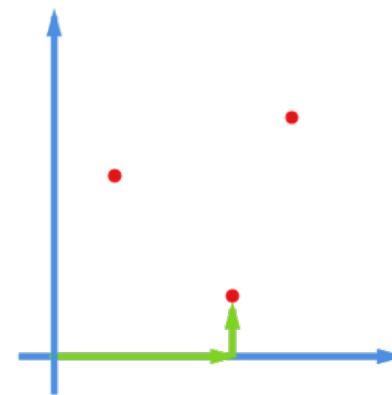
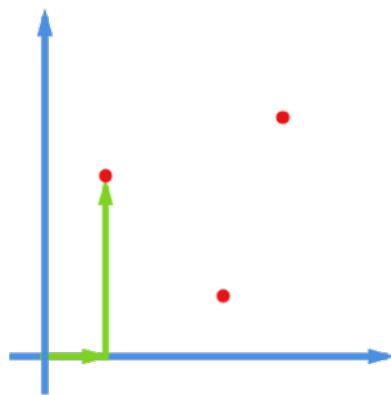
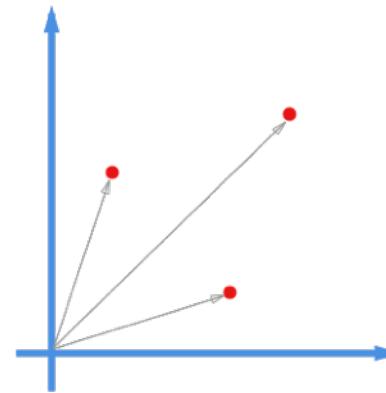
### “Primera” descripción del problema

Dado un registro incompleto de valoraciones de las películas vistas por los usuarios, nos gustaría saber cuáles películas no vistas por los usuarios son mas probables que cada usuario vea.

## Abstracción del problema: $V$

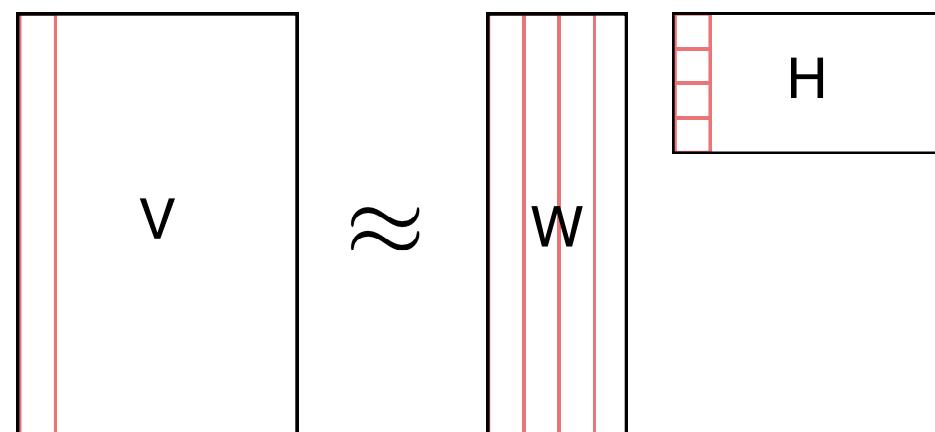
$$V =$$

4	4
1	3
3	1



## Transformación del problema

$$V \approx WH^T$$



$$\mathbf{v}_j \approx \sum_k h_{kj} \mathbf{w}_k$$

## “Segunda” descripción del problema (formular el problema + imponer restricciones)

Dada una matriz  $V_{m \times n}$  con entradas faltantes, encontrar la factorización de menor rango, con entradas no negativas, que aproxime la matriz original,  $V_{m \times n} \approx W_{m \times k} H_{k \times n}^T$

$$\begin{aligned} \min_{W,H} \quad & \|V - WH^T\|_F^2 \\ \text{s. t. } & W, H \geq 0 \end{aligned}$$

## Proponer una solución

---

### Algorithm 1 NMF con datos incompletos

---

- 1: **Input**  $V$ ,  $k_0$ ,  $\Omega = \{\text{entradas NAN}\}$
- 2: **Initialize**  $[Z = \{V\}, Z_\Omega = \{0\}]$ ,  $W \sim U(0, 1)$ ,  $H \sim U(0, 1)$
- 3: **Output**  $W, H$
- 4: **repeat**

$$\begin{aligned}H &\leftarrow [Z^T W] [W^T W]^{-1} \\W &\leftarrow [ZH] [H^T H]^{-1} \\Z_\Omega &\leftarrow [WH]_\Omega\end{aligned}$$

- 5: **until** convergence
  - 6: **Return**  $W, H$
-

## Implementar solución (ver código)

```
def addmm_nmfcc(V_inc, W0, H0, rho, max_iter, omega):
    #Dimensions
    (m,n) = V_inc.shape
    k = W0.shape[1]

    #Initialize
    Z = np.copy(V_inc)
    Z[omega] = 0.0
    W = np.copy(W0)
    H = np.copy(H0)
    W_plus = np.copy(W)
    H_plus = np.copy(H)
    alpha_W = np.zeros_like(W)
    alpha_H = np.zeros_like(H)

    for i in range(max_iter):
        #Compute H
        H = np.dot(np.linalg.pinv( np.dot( np.transpose(W), W ) + rho*np.eye(k) ), np.dot( np.transpose(W), Z ) + rho

        #Compute W and condition number of (HHt+I)
        W = np.transpose( np.dot( np.linalg.pinv( np.dot(H,np.transpose(H)) + rho*np.eye(k) ), np.dot(H, np.transpose(Z

        #Compute Z
        Z[omega] = np.dot(W,H)[omega]

        #Compute W_plus
        W_plus = np.clip( W + (1.0/rho)*alpha_W, 0, np.inf )

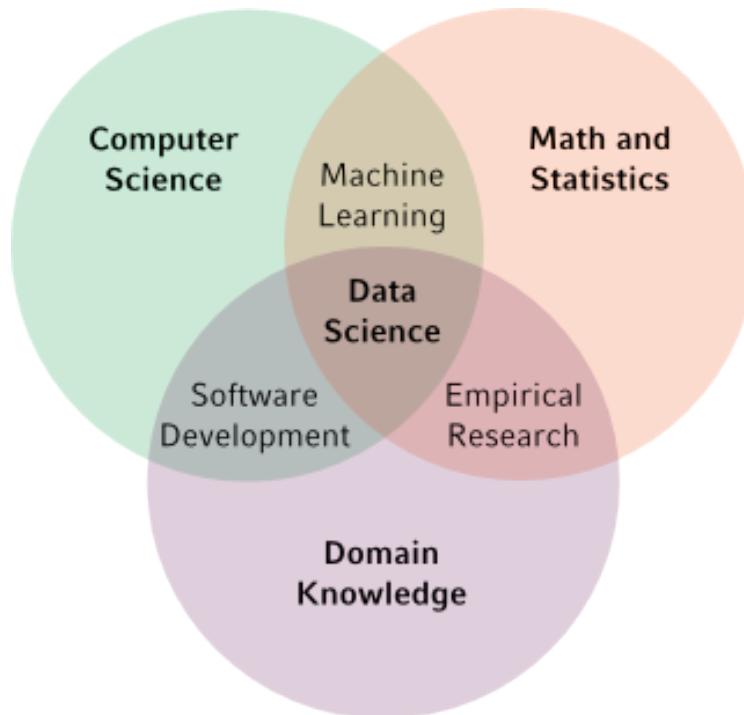
        #Compute H_plus
        H_plus = np.clip( H + (1.0/rho)*alpha_H, 0, np.inf )

        #Compute alpha_W
        alpha_W = alpha_W + rho*W - rho*W_plus

        #Compute alpha_H
        alpha_H = alpha_H + rho*H - rho*H_plus

    return (W_plus, H_plus)
```

# Científico de datos



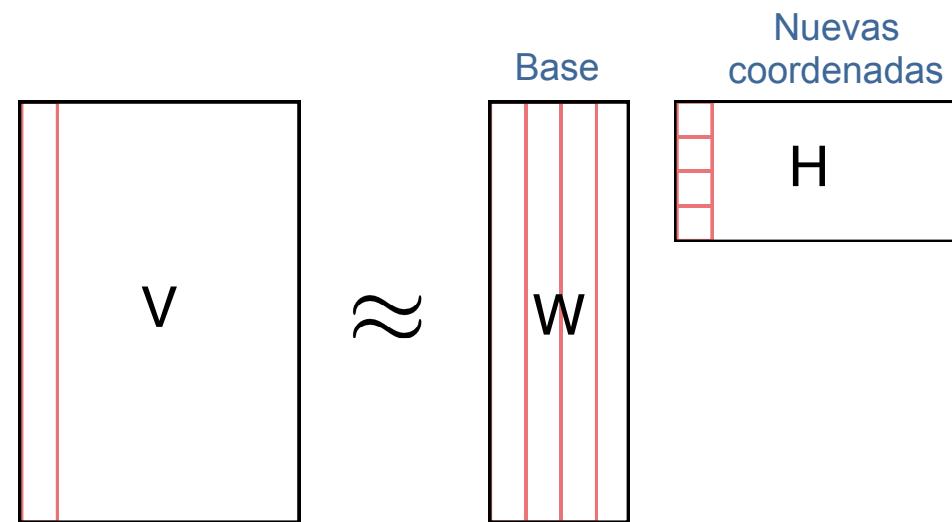
# Conexión con otros problemas (Factorización de matrices no negativas)

# Factorización de Matrices no Negativas (NMF)

Método de **reducción de dimensionalidad** propuesto por Lee y Seung [1]

$$\min_{W,H} F(V | WH^T)$$

$$\text{s. t. } \begin{aligned} W &\geq 0 \\ H &\geq 0 \end{aligned}$$

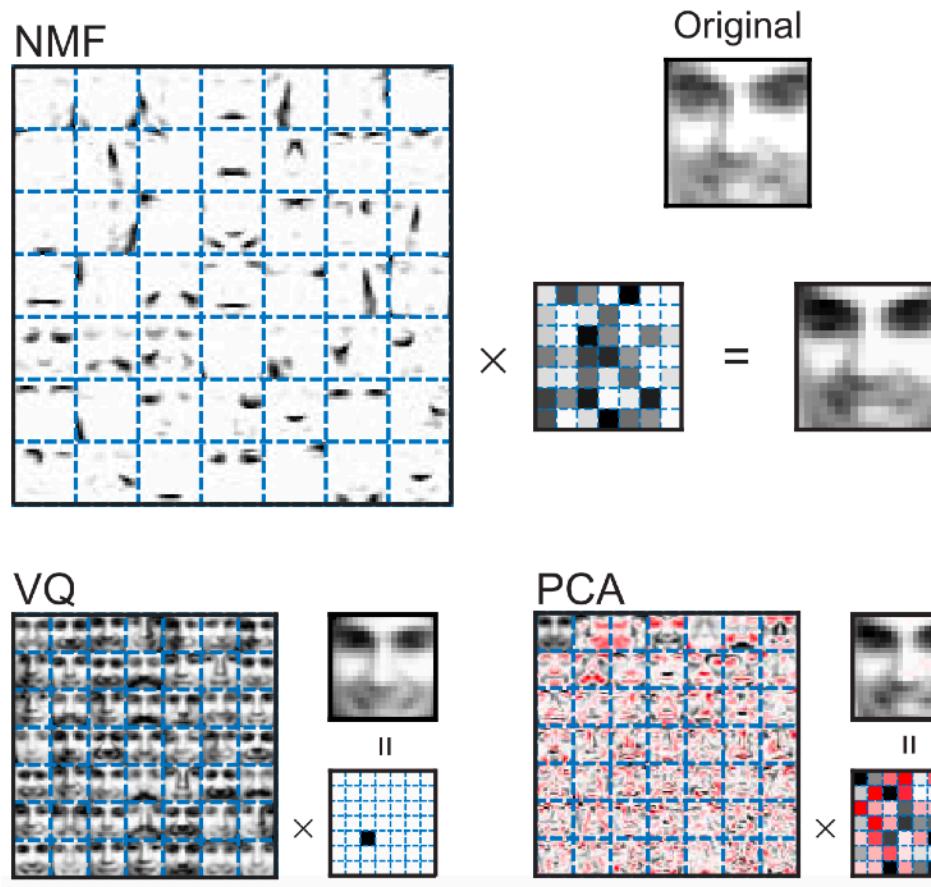


$$\mathbf{v}_j \approx \sum_k h_{kj} \mathbf{w}_k$$

La restricción de no negatividad promueve una **representación de los datos basada en partes**, lo cual permite “sumar” partes, no restar.

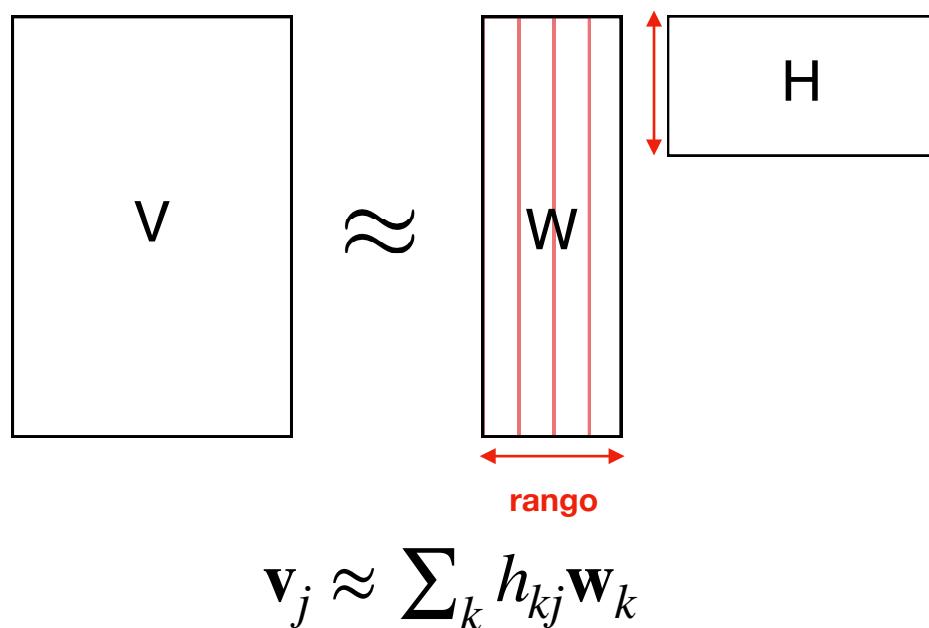
# Factorización de Matrices no Negativas (NMF)

Diferentes métodos de reducción de dimensionalidad: NMF, PCA, VQ [1]



# Factorización de Matrices no Negativas (NMF)

Con la correcta elección del rango, NMF produce de una forma natural una representación de los datos basada en partes.

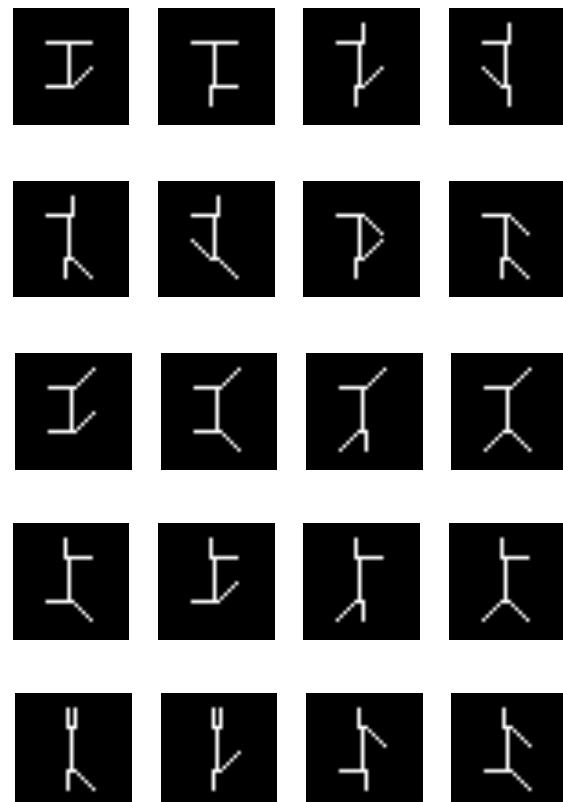


El **rango** en NMF tiene un **significado real (interpretación)**: determina el número de rasgos que se extraerán de los datos [1]

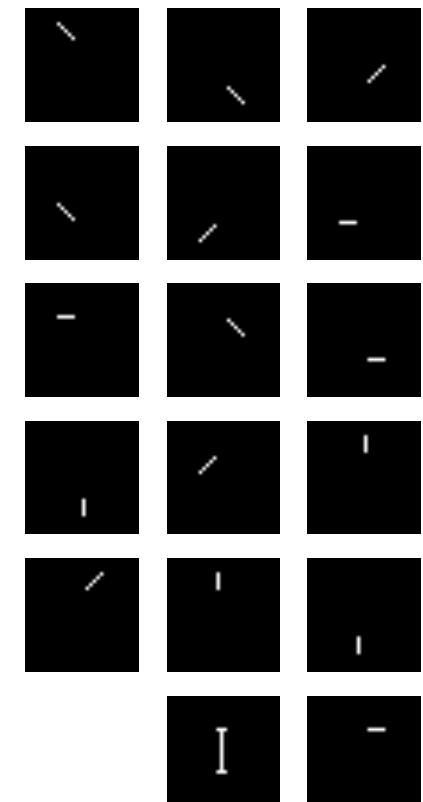
# Factorización de Matrices no Negativas (NMF)

## NMF en imágenes

Swimmer data set [1]



base W



# Factorización de Matrices no Negativas (NMF)

## NMF en otros casos

NMF en textos



base W

Tópicos

NMF en ratings

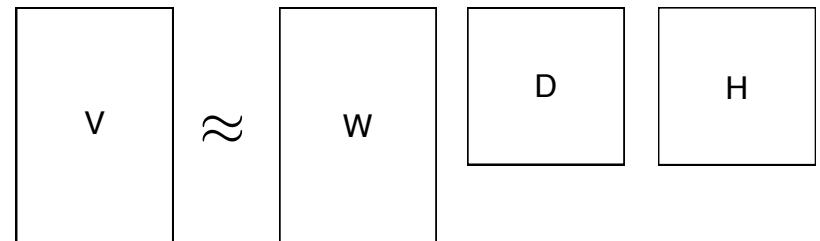


“Estereotipos”

## Cómo resolver el problema de estimar el rango óptimo

Minimizar el rango a través de un término de regularización

$$\begin{array}{ll}\min_{W,D,H} & F(V|WDH^T) + \text{Tr}(D) \\ \text{s. t.} & W \geq 0, \quad \|w_k\|_2 = 1 \quad \forall k \\ & H \geq 0, \quad \|h_k\|_2 = 1 \quad \forall k \\ & D \geq 0.\end{array} \quad (1)$$

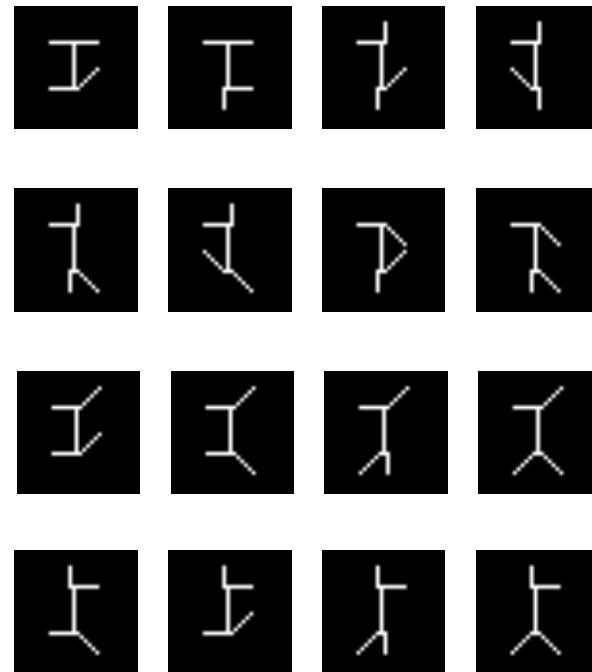


Penalizar los elementos de  $D$  ayuda en minimizar el rango en NMF ?

# Experimento 1: NMF imágenes

## NMF en imágenes

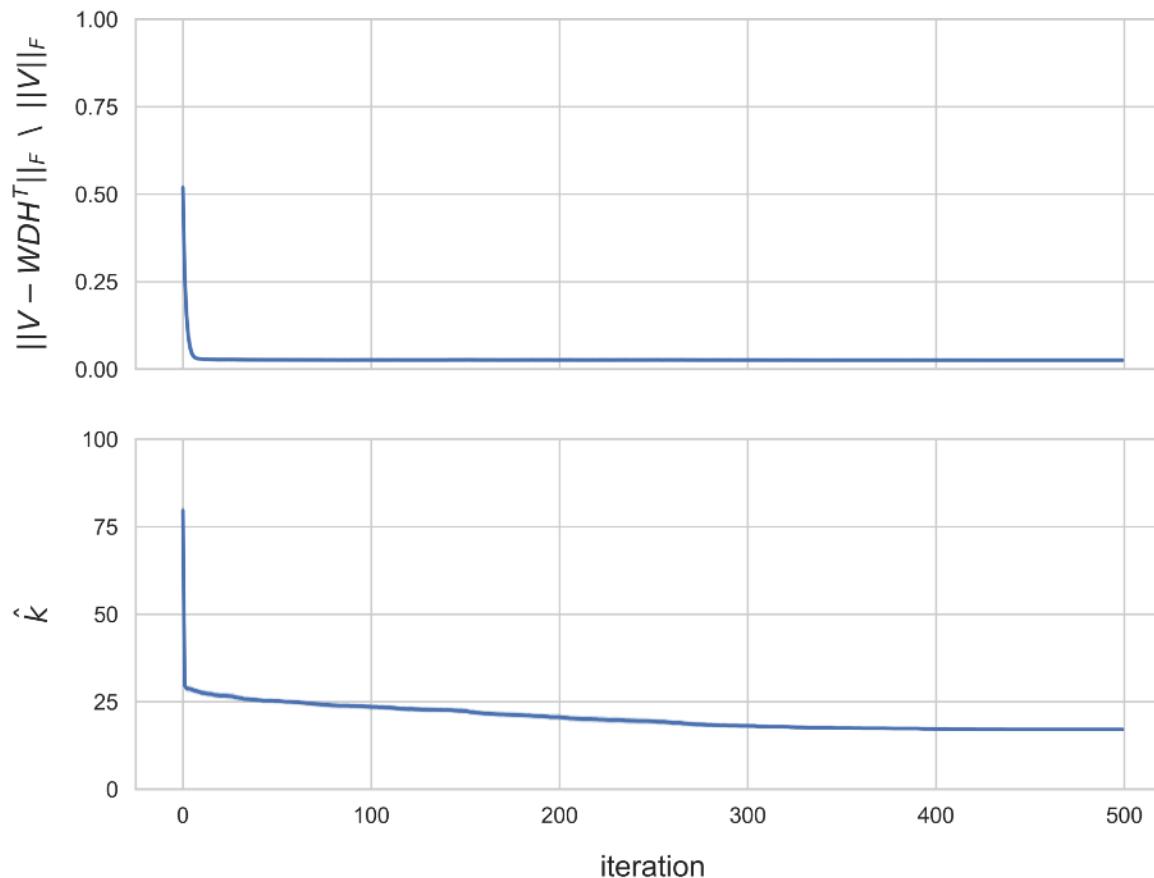
El conjunto de datos Swimmer [1] contiene 256 imágenes en escala de grises con un resolución de  $32 \times 32$ . Cada imagen muestra una figura con un torso estático y cuatro extremidades móviles.



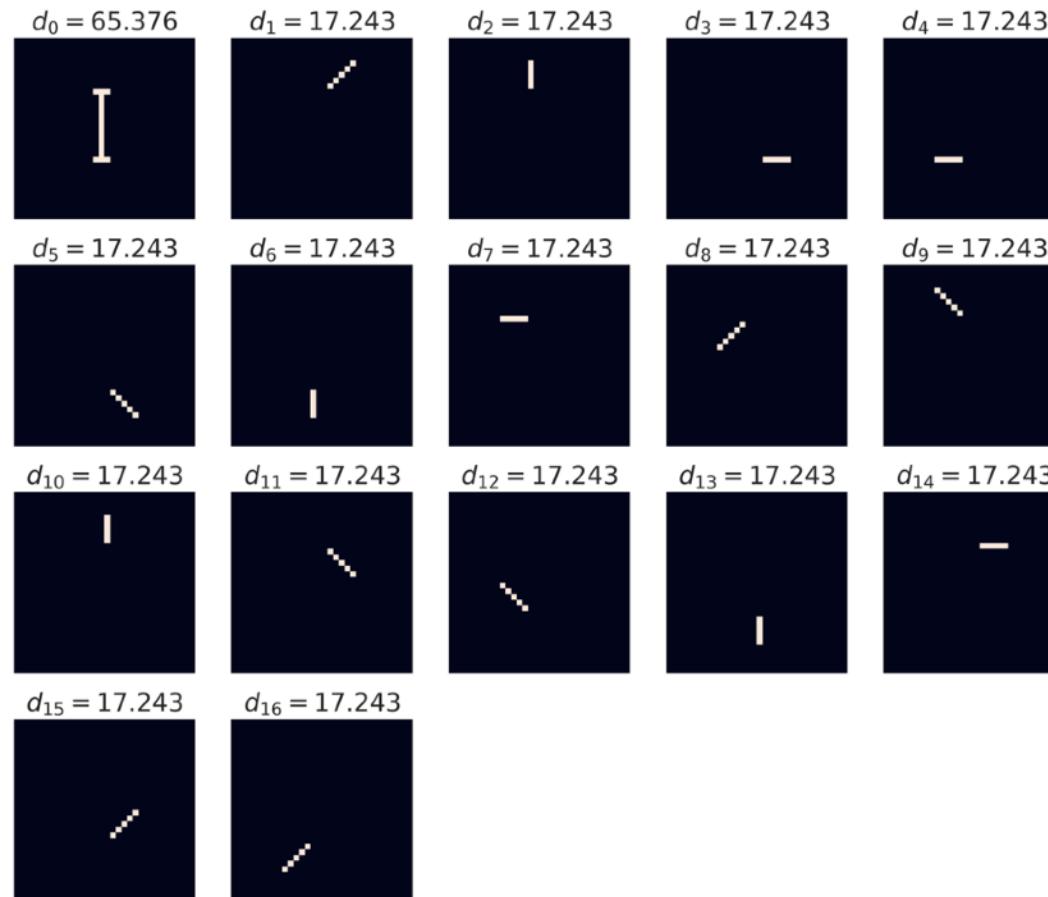
Las 256 imágenes se apilan por columnas para formar una matriz con dimensiones  $1024 \times 256$

# NMF en imágenes

Para observar el **comportamiento del algoritmo durante las iteraciones**, se realizaron 30 ejecuciones independientes con los parámetros fijos  $k_0 = \min(1024, 256)$  y  $(\alpha_W, \alpha_H, \alpha_D, \theta) = (0.01, 0.01, 0.6, 0.8)$

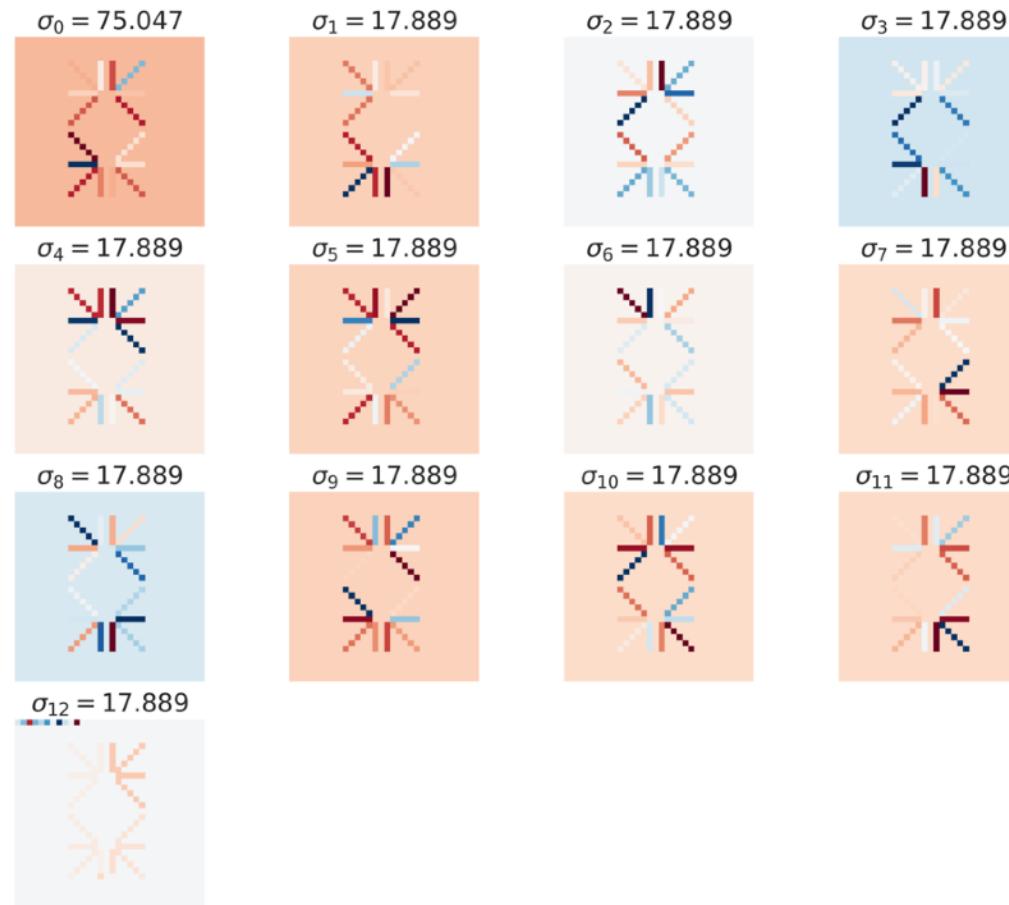


# NMF en imágenes



# NMF en imágenes

Comparación con SVD, éste estima un rango de 13 elementos. Como referencia, se muestran los 13 *eigen-parts*



# Experimento 2: NMF videos

## *Background subtraction*



# **Experimento 3: NMF textos**

## Topic modeling

El conjunto de datos 20 Newsgroups [1], contiene 20000 documentos distribuido en 6 grandes tópicos, a su vez distribuidos en 20 sub-tópicos.

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

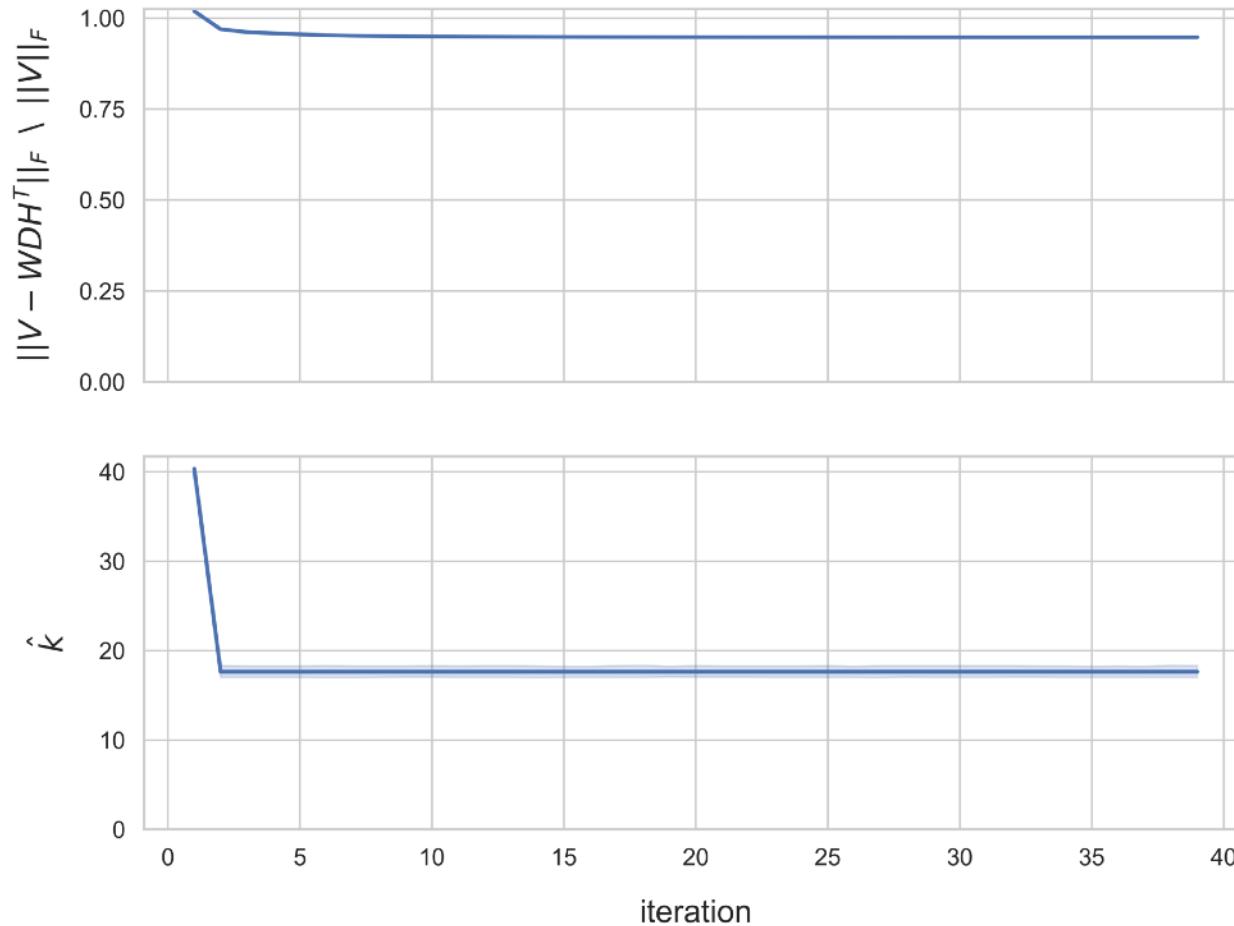
Se seleccionan 2000 documentos aleatoriamente y 1000 características de dichos documentos con los valores más altos de la métrica term *frequency – inverse document frequency* (tf-idf).

Cada documento se apilan por columnas para formar una matriz con dimensiones  $1000 \times 2000$ .

[1] K. Lang, “Newsweeder: Learning to filter netnews” in Proceedings of the Twelfth International Conference on Machine Learning, 1995, pp. 331-339.

# NMF en textos

## Comportamiento del algoritmo durante las iteraciones



# NMF en textos

## Topic #0:

scsi controller dec card error disk various bit result message

## Topic #1:

israel israeli jews attacks soldiers state allow policy true center

## Topic #2:

game games team season year win time bad goal play

## Topic #3:

00 sale 10 today condition new equipment card price 20

## Topic #4:

software available public developed specifically systems computer appreciated thanks difference

## Topic #5:

windows dos file use os program ms using files unix

## Topic #6:

car cars new miles good tires insurance like year used

## Topic #7:

people think don just like know government make say good

# NMF en textos

## Topic #8:

drive drives disk hard floppy mac computer mb original software

## Topic #9:

chip clipper used encryption want product follow don secure years

## Topic #10:

card driver thanks drivers video does anybody cards ram ll

## Topic #11:

thanks advance know hi does unix info mail interested anybody

## Topic #12:

window manager application request problem use received right exactly information

## Topic #13:

bike insurance good know live course recommend contact don work

## Topic #14:

space nasa sci launch orbit files station ll know old

## Topic #15:

key keys chip public encryption use government clipper communications message

## Topic #16:

god jesus bible faith does people believe christ heaven life

# Experimento 4: NMF genes

# NMF en genes

El conjunto de datos Leukemia [1] [2] consta de **5000 mediciones** de expresión genética de **38 muestras** de médula ósea. Las 38 muestras se dividen en las siguientes clases:

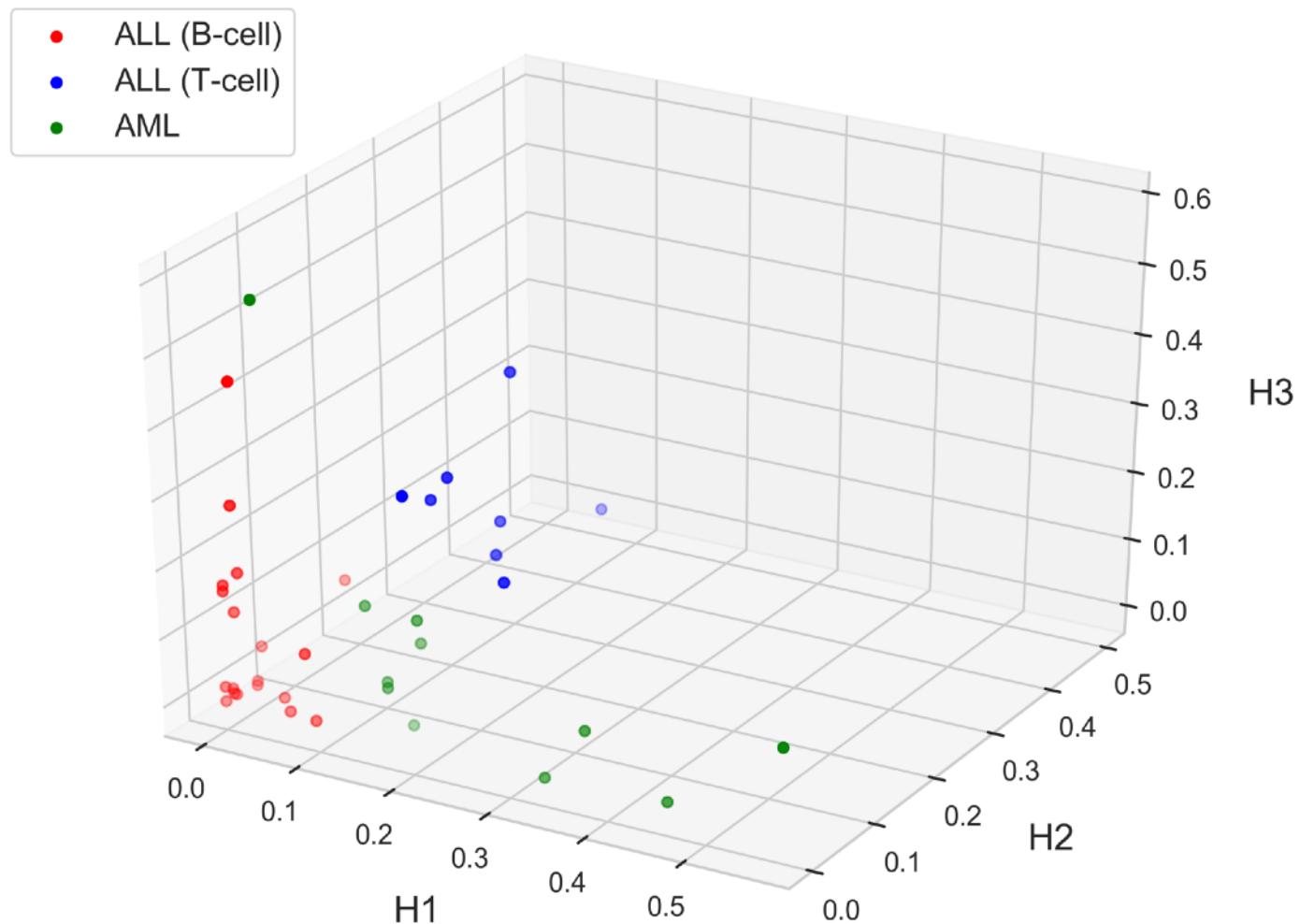
- 11 *Acute Myelogenous Leukemia* (AML)
- 27 *Acute Lymphoblastic Leukemia* (ALL)
  - ▶ 19 de la subclase con células tipo B
  - ▶ 8 de la subclase con celular tipo T

Generalmente se asume un rango entre 2 y 5 en aplicaciones de NMF [3].

- [1] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, “Metagenes and molecular pattern discovery using matrix factorization,” Proc. Natl. Acad. Sci. USA, vol. 101, no. 12, pp. 4164–4169, 2004.
- [2] K. Devarajan, “Nonnegative matrix factorization: An analytical and interpretive tool in computational biology,” *PLoS Comput. Biol.*, vol. 4, no. 7, 2008.
- [3] S. Squires, A. Prügel-Bennett, and M. Niranjan, “Rank Selection in NonnegativeMatrix Factorization using Minimum Description Length,” *Neural Comput.*, vol. 29, no. 8, pp. 2164–2176, 2017.

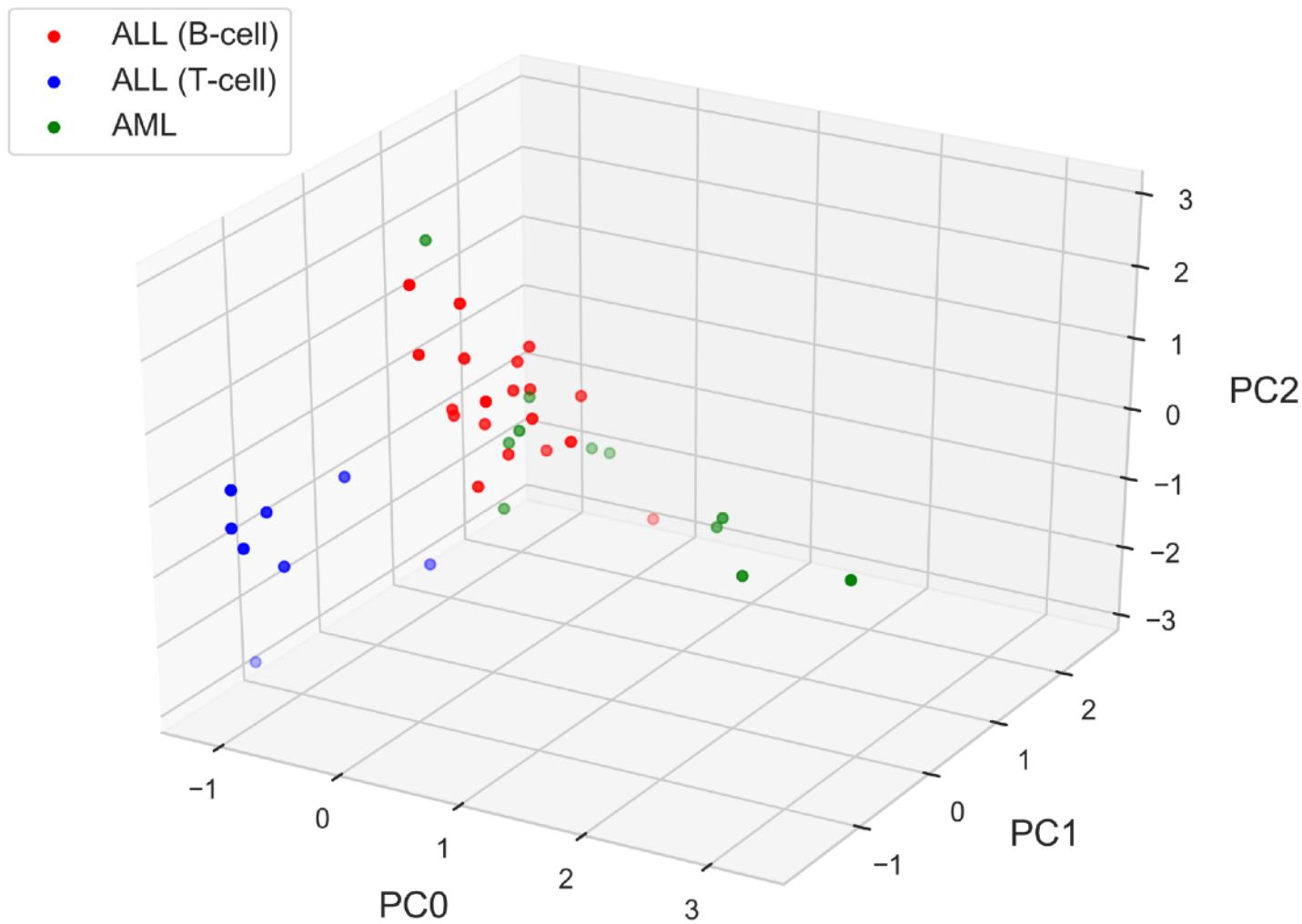
# NMF en genes

Proyección estimando el rango óptimo



# NMF en genes

Proyección usando 3 primeros componentes (PCA)



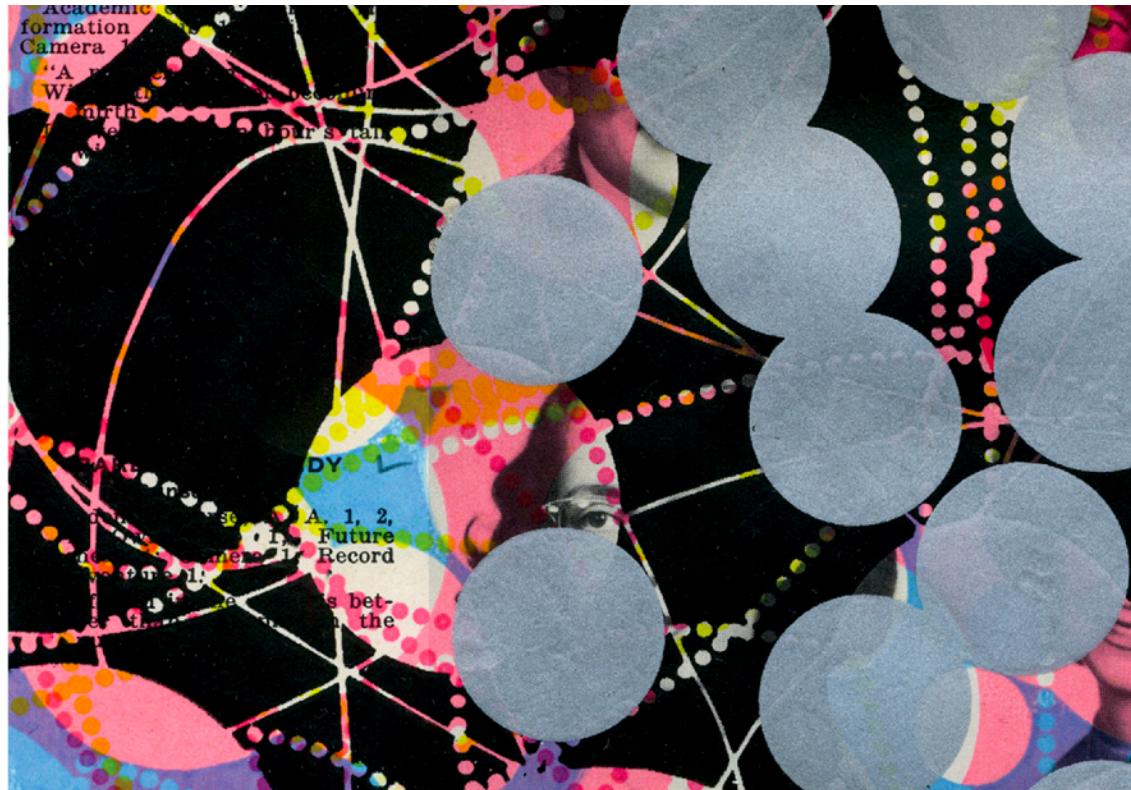
# Factorización de Matrices no Negativas con regularización de rango

**La correcta elección del rango permite extraer rasgos localizados de los datos.**

**Nuestra propuesta estimó correctamente el rango tanto datos sintéticos como en imágenes, genes y textos.**

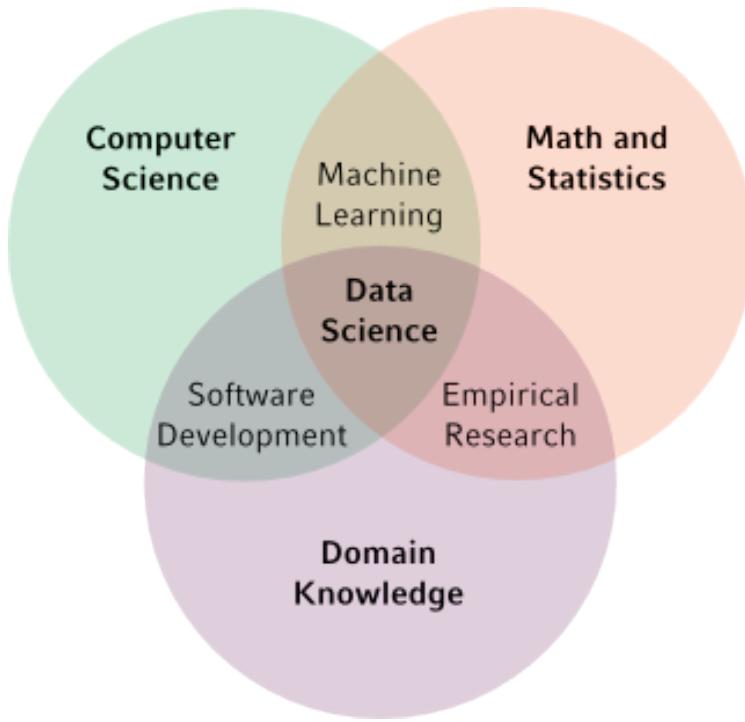
**Nuestra propuesta puede adaptarse en diversas formulaciones exitosas de NMF: HALS, ADMM-NMF, etc.**

# Ideas Finales



DATA

# Data Scientist: The Sexiest Job of the 21st Century



Conocimiento y  
Creatividad

Estrategia -  
Plan

## **Odin Eufracio**

Centro de Investigación en Matemáticas - CIMAT  
Jalisco SN, Mineral de Valenciana Gto. Gto.

Office: D307

Phone: (+52) 473 732 7155 ext. 4730

E-Mail: [odin.eufracio@cimat.mx](mailto:odin.eufracio@cimat.mx)