Reducción de dimensionalidad

Fisica computacional 2

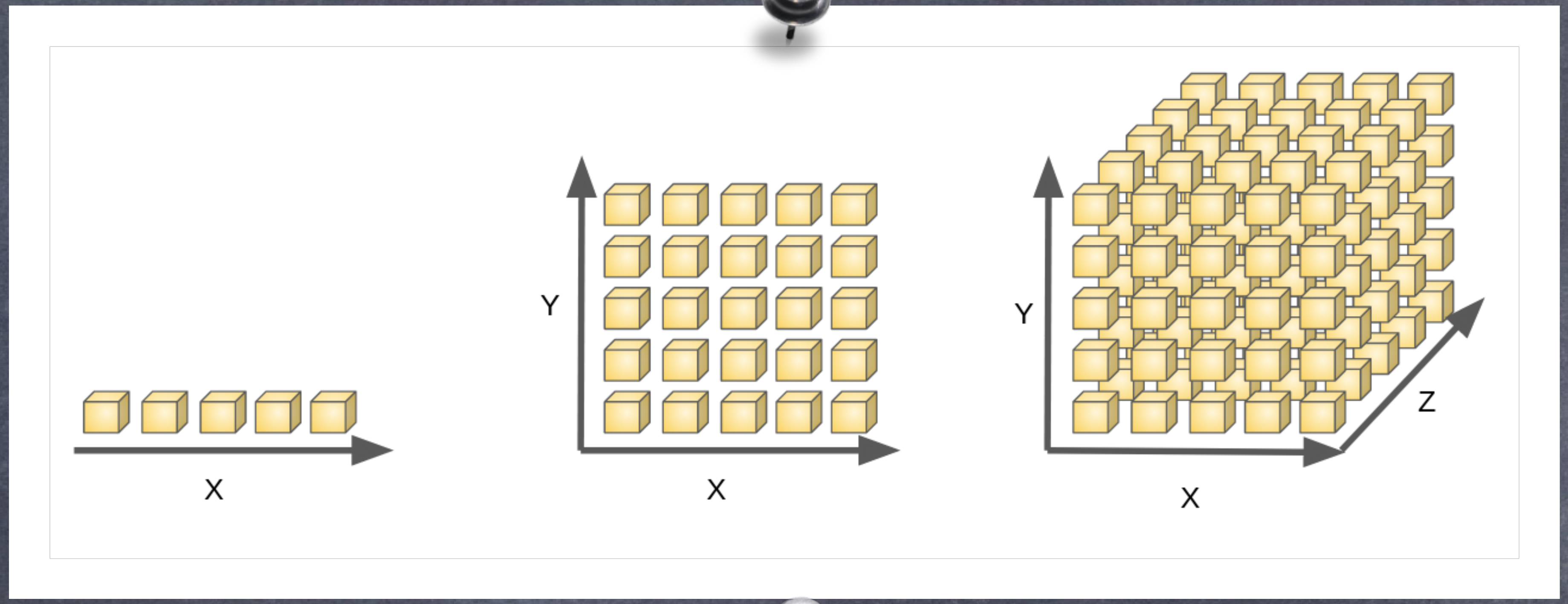
Ph.D. Santiago Echeverri Arteaga

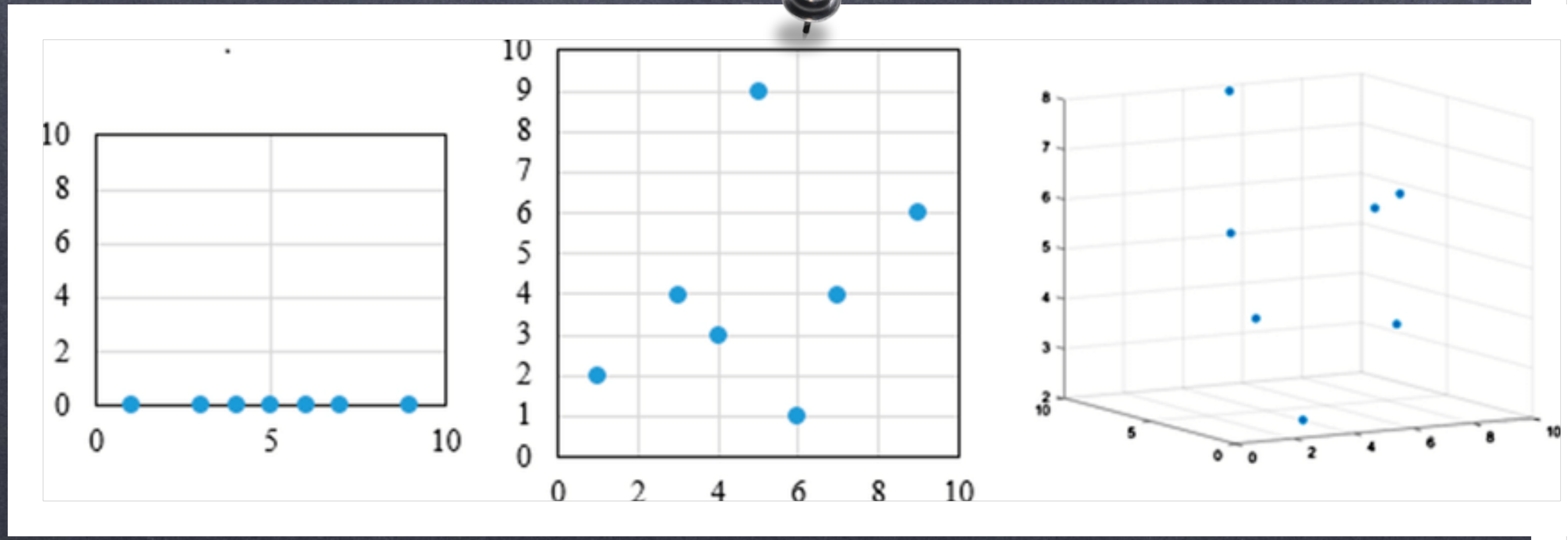
¿Que es la dimensión?

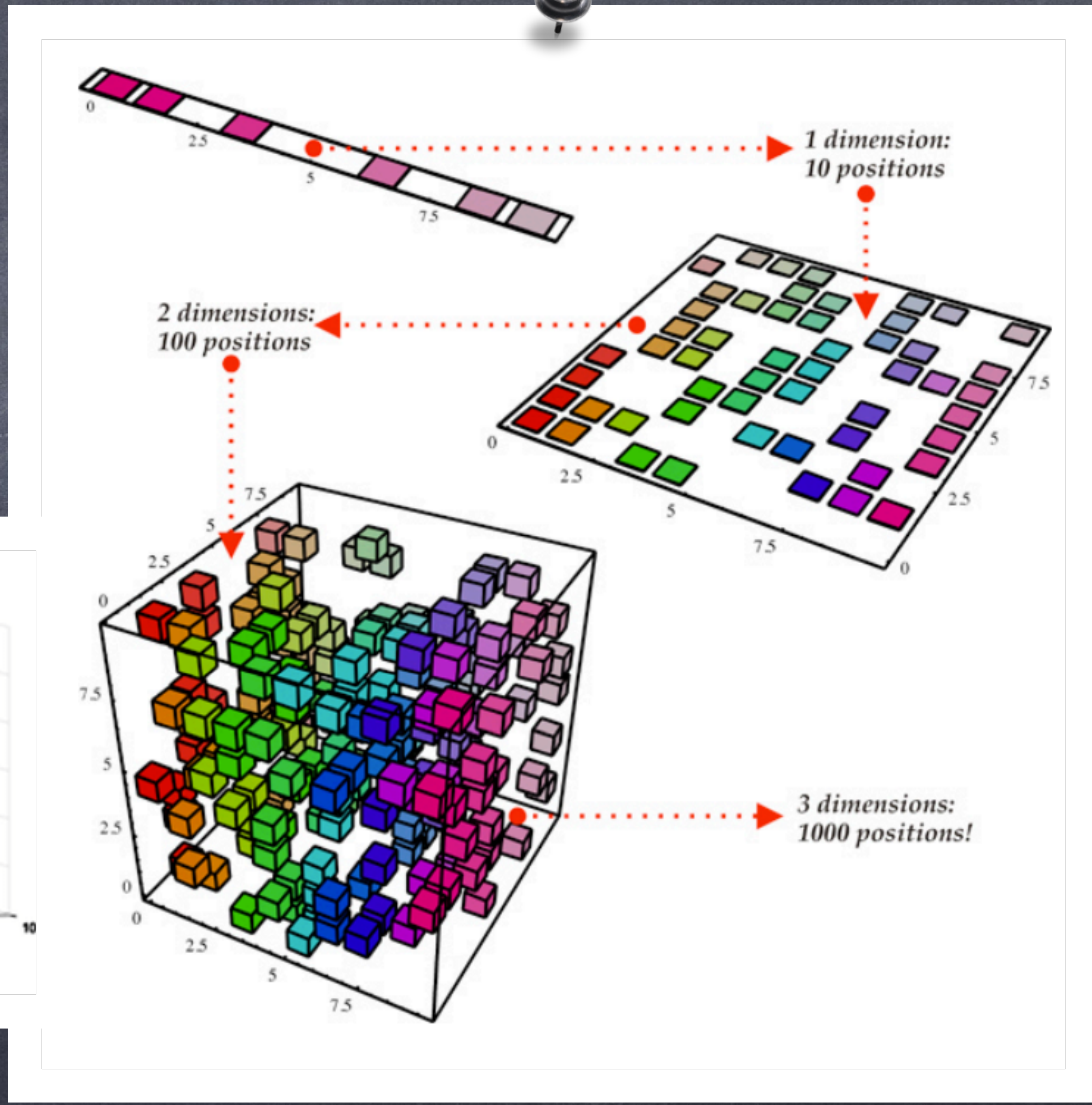
- Espacios vectoriales: Número de elementos de la base del sub-espacio vectorial y por tanto nos habla de los grados de libertad. EJ: Persona ante diferente iluminación (Sub-EV cerrando ante la multiplicación)
- PROBLEMA: Círculo en $\mathbb{R}^3$ no es bien comportado ante combinaciones lineales → Definir distancia desde las variedades. Círculo puede ser rapeado a una línea

Curse of dimensionality

- Datos dispersos
- Cantidad de datos necesarios para entrenar un modelo crece exponencialmente
- Volumen crece exponencialente







¿Qué es la dimensión?

- Dimensión Fractal
- Dimensión de un conjunto de imágenes (Tensores de rango 3) que rotadlas y con diferente iluminación. Grados de libertad asociados
- ¿Cómo afecta el ruido a la dimensionalidad?
- Reducción de dimensionalidad: Preservar cierta cantidad de interés al hacer un rapeo hacia un espacio de dimensionalidad inferior (Como pairwise distance), la similaridad entre puntos, o la estructura local

Análisis de componentes principales (PCA)

- Sea $x^i \in \mathbb{R}^m$, se puede descomponer en términos de la suma entre la proyección sobre el vector $a = Px$ y el residuo ortogonal $b = (I - P)x$ así: $x = Px + (I - P)x$ con $P = UU^T$ y $U^TU = I$

- Se tiene que $\|x\| = \|a\| + \|b\|$

- Definiendo la matriz de datos como $X = [x^1, \ldots, x^n]$ se tiene que $\|X\| = \sum_{i=1}^{n} x^i = \|A\| + \|B\|$

- Se desea maximizar la suma de las distancias proyectadas $\max_{U^TU=I} \|UU^TX\|$

- Usando multiplicadores de Lagrange se obtiene $XX^Tu_i = \lambda_i u_i$

Análisis de componentes principales (PCA)

- Y como XX^T es simétrica, definida positivamente (Distancia) se pueden ordenar los autovalores ($\lambda_1 \geq \ldots \geq \lambda_n$)
- Si a los datos se les ha substraído la media (centrados en cero) XX^T cada autovalor representa la varianza en esa dirección particular (La que dictamina el autovector)
- El espectro de autovalores indica la distribución de información del espacio vectorial ($-\sum_i \lambda_i \ln \lambda_i$)
- Si todos los autovalores son iguales, la entropía es máxima → No hay direcciones privilegiadas en el set de datos y por tanto no se puede reducir su dimensionalidad. Si un solo autovalor es diferente de cero, el set de datos se puede reducir a una dimension
- La base de PCA es aquella que minimiza la entropía de Shanon

PCA

- Las siguientes afirmaciones sobre PCA son equivalentes
  - Maximiza la longitud cuadrada de los datos proyectados,
  - Minimiza la longitud cuadrada de los residuos,
  - Maximiza la varianza estadística de los datos reducidos
  - Minimiza la entropía de Shannon, y produce coordenadas en las que los datos reducidos de dimensión no se repiten.
- Las direcciones con autovalores más bajos corresponderán al ruido
- Mantiene la relación entre las distancias
- PCA es lineal pero se puede aplicar un Kernel previo para hacerlo no-lineal

Conexión con SVD: Descomposición en valores singulares

- Cada matriz rectangular X se puede descomponer en un producto de matrices ortogonales U, V, y una matriz diagonal Σ. Las matrices U, V contienen los auto-vectores singulares izquierdos y derechos respectivamente y Σ los valores singulares ordenados. $X = U\Sigma V^T$

- Los auto-vectores izquierdos son los auto-vectores que surgen de PCA: $XX^T = U\Sigma^2 U^T$

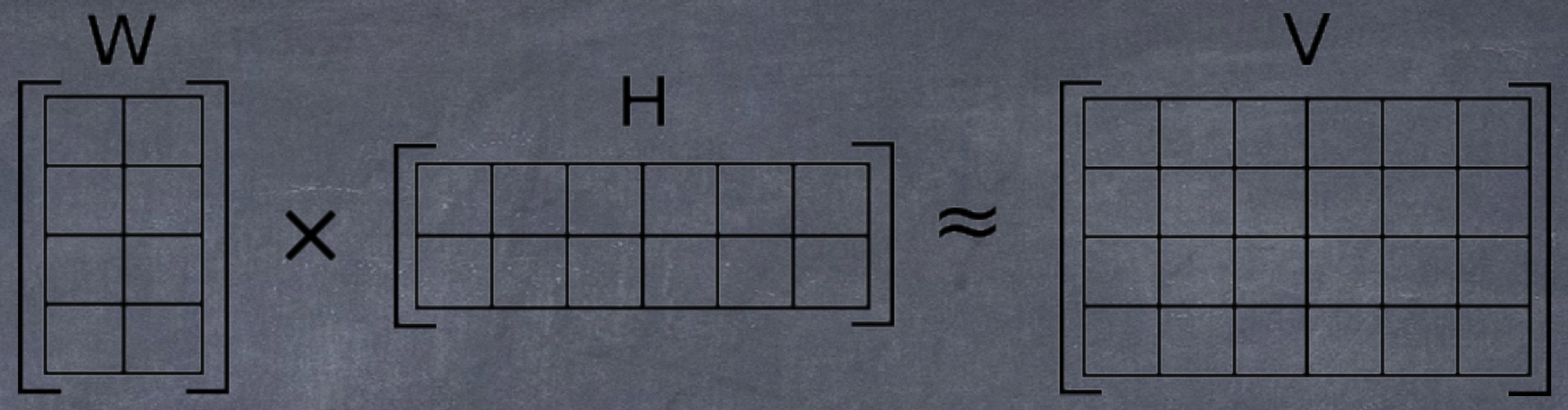
- SVD Truncada a k dimensiones: $X_{m\times n} \approx U_{m\times k}\Sigma_{k\times k}V^T_{k\times n}$

- Se toman la cantidad de dimensiones que expliquen el porcentaje de varianza deseado

Escalado multidimensional

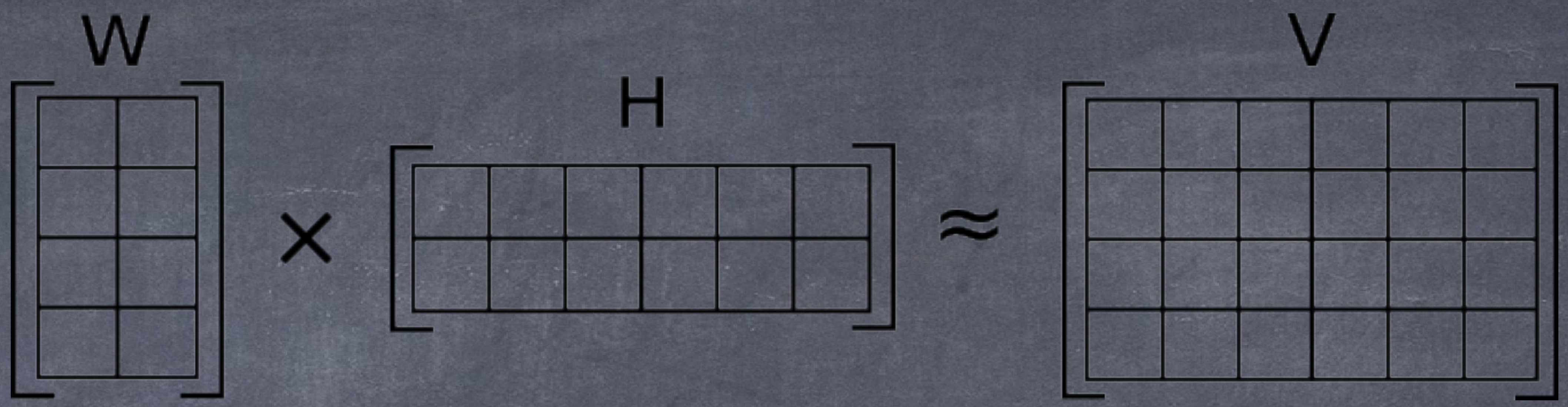
- ¿Cómo encontrar un espacio donde las distancias euclidianas sean lo más fieles posible a la matriz de distancia proporcionada (o calculada)?
- Se calcula la matriz de distancias (si se tiene la matriz de norma/similaridad S_{ij} se toma $d_{ij} = S_{ii} - S_{jj} - 2ij$)
- Se encuentran los z que minimizan la tensión $\sum_{i \neq j = 1}^{N} \left((d_{ij} - |z_i - z_j|)^2 \right)^{\frac{1}{2}}$
- Variantes:
  - Isomap: KNN para mantener el ordenamiento
  - TSNE: Puntos cercanos/lejanos permanecen cerca/lejos

Factorización no negativa de matrices

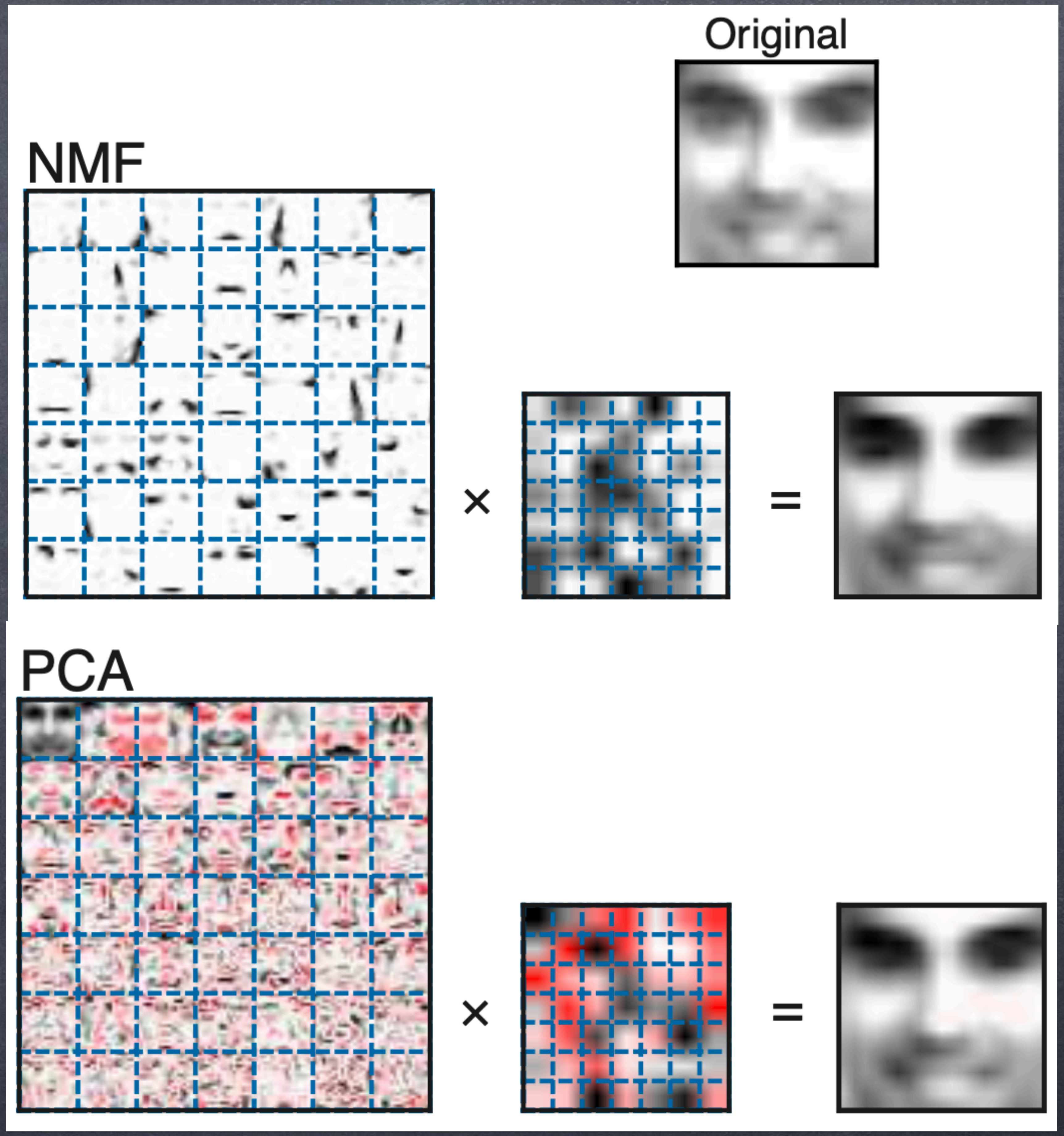


- Usado para tensores no negativos. EL objetivo es encontrar $W, H : X \approx WH$ donde X es $N \times p$, W es $N \times r$ y H es $r \times p$. Además $r \leq \max(N, p)$
- Se asume $x_{ij}, w_{ij}, h_{ij} \geq 0$
- Log Likelihood: $\sum_{i=1}^{N} \sum_{j=1}^{p} [x_{ij} \log(WH)_{ij} - (WH)_{ij}]$ en donde x_{ij} tiene una distribución de Poisson centrada en $(WH)_{ij}$
- No tiene solución única
- Matriz W Términos → Tópicos: Qué tanto cada base de H representa a X
- Matriz H Tópicos → Documentos: X es una combinación lineal de las filas de H
- Como todos los componentes son positivos su interpretación es más directa

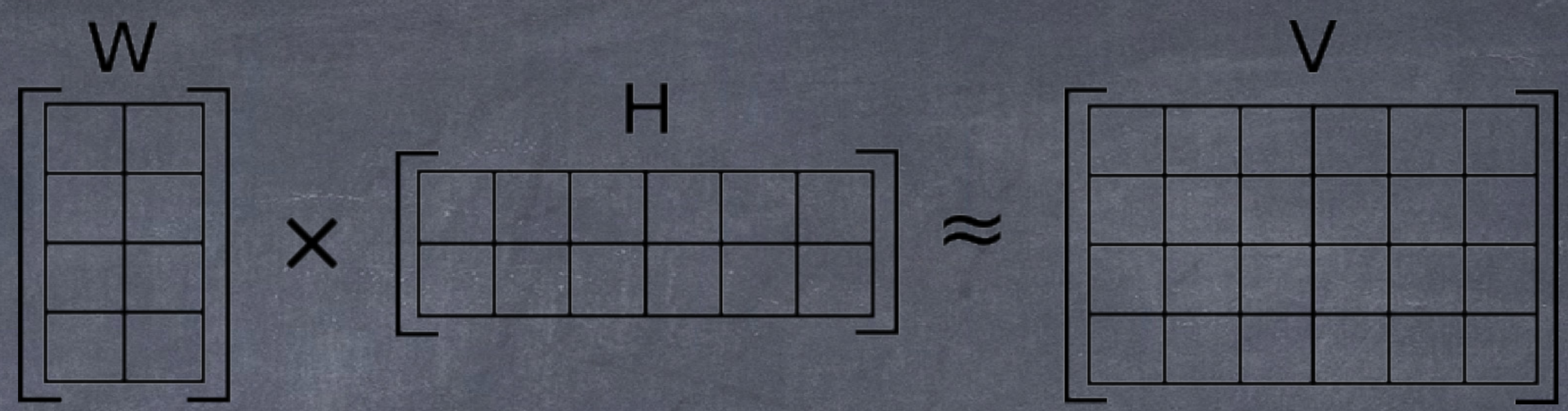
Factorización no negativa de matrices



- Usado en procesamiento de lenguaje natural y reconocimiento de lenguaje:
  - Filas: Docuementos
  - Columnas: Palabras
  - Valores: Conteo de palabras
- Usado para dividir una imagen en sus componentes
- Minería de texto
- Encriptación
- Video / Música / Imágenes



Factorización no negativa de matrices



- Usado en procesamiento de lenguaje natural sobre matrices td-if
- Matriz de frecuencia de términos tf: Que tanto aparece cada palabra en cada documento
- Matriz frecuencia de término-frecuencia de documento inversa td-tif:
  - $idf = \log\left(\frac{N}{|d \in D : t \in D|}\right) + 1$ Reduce la importancia de las palabras que aparecen mucho en todos los documentos
  - $|d \in D : t \in D|$: Número de documentos donde el término t aparece.
  - $tf - idf = f_{td} * idf$
  - f_{td}: Frecuencia de aparición de la palabra t en el documento d

Uso de métodos

Método	USO
PCA	Identificar un número pequeño de variables manteniendo la varianza
Kernel PCA	Relaciones no lineales
Escalado Multidimensional	Como PCA pero manteniendo distancia entre puntos. Visualizar clusters
Factorización no negativa de matrices	Solo se tienen valores positivos (Imágenes o palabras)