Bases de Datos

Recuperación de Información

Lecturas

https://es.wikipedia.org/wiki/B%C3%BAsqueda_y_recuperaci%C3%B3n_de_informaci%C3%B3n

¿Qué es Recuperación de Información?

¿Qué es Recuperación de Información?

¿Qué es Recuperación de Información?

¿Qué es Recuperación de Información?

satisfacer una necesidad de información encontrando material (generalmente documentos) desestructurado (generalmente texto, pero también imágenes, sonido, video) albergado en grandes colecciones

¿Qué es Recuperación de Información?

satisfacer una necesidad de información encontrando material (generalmente documentos) desestructurado (generalmente texto, pero también imágenes, sonido, video) alber

distinto de question answering o satisfacer queries

¿Qué es Recuperación de Información?

satisfacer una necesidad de información encontrando material (generalmente documentos) desestructurado (generalmente texto, pero también imágenes, sonido, video)

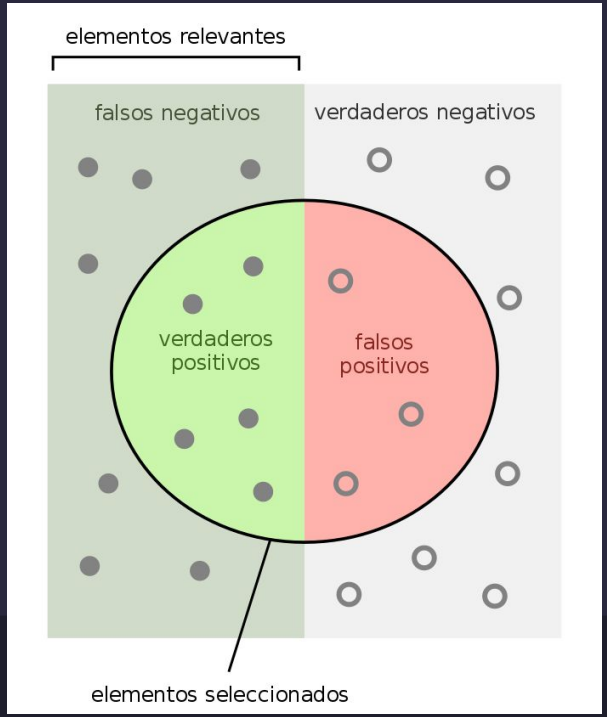
resulta crítico el modelado cognitivo, comunicativo de la consulta

Definir el éxito de una consulta

No es binario, sino basado en métricas

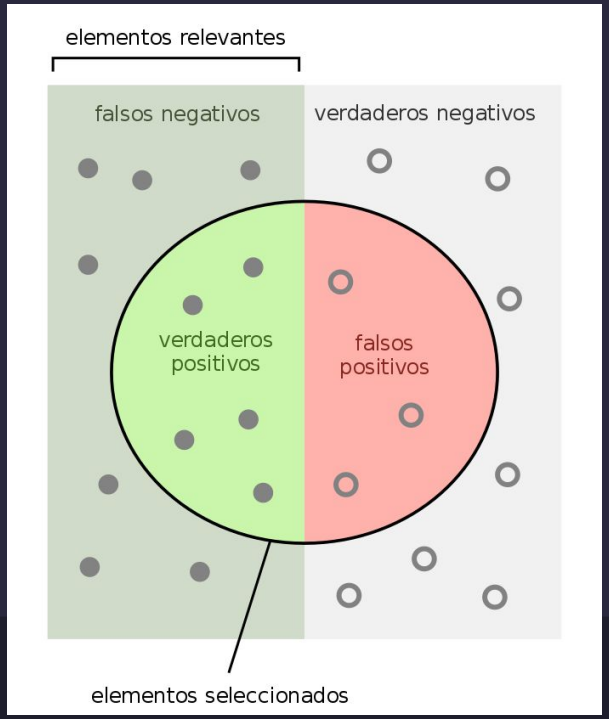
Definir el éxito de una consulta

No es binario, sino basado en métricas



Definir el éxito de una consulta

No es binario, sino basado en métricas



¿Cuántos objetos relevantes se seleccionaron?
i.e. Cuantas personas enfermas son identificadas como tales.

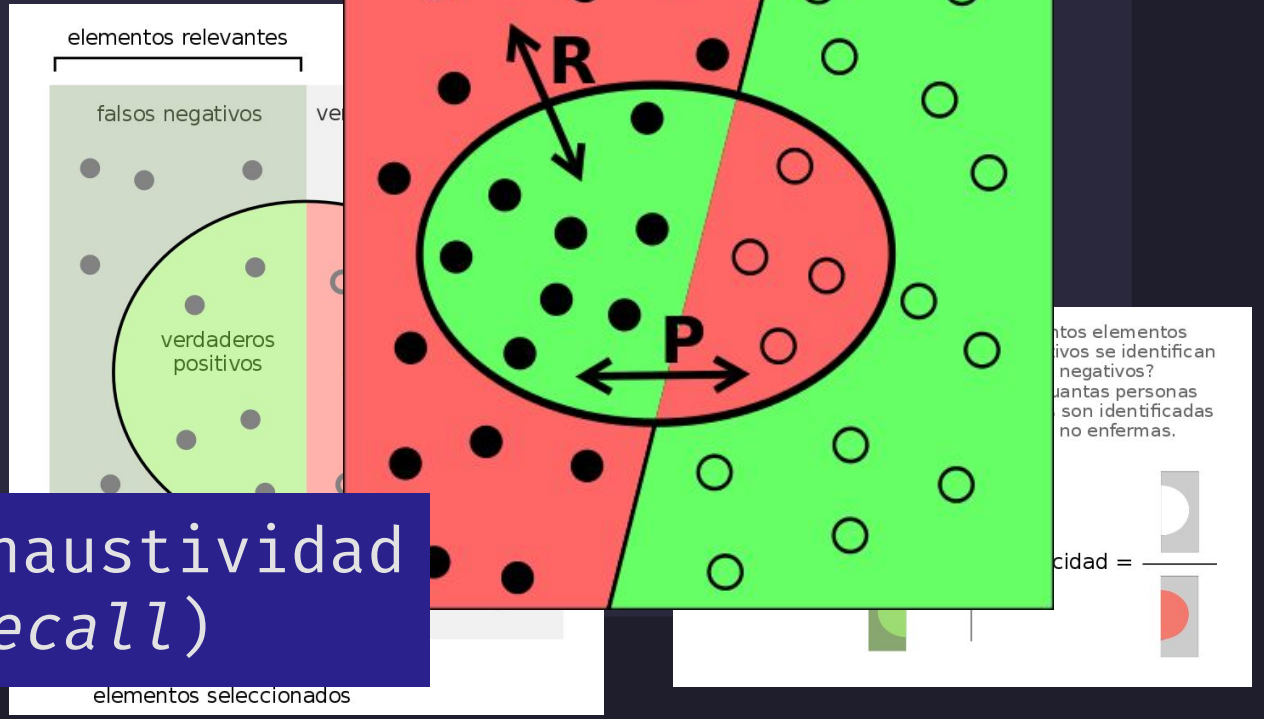
Sensibilidad = $\frac{\text{verdaderos positivos}}{\text{elementos relevantes}}$

¿Cuántos elementos negativos se identifican como negativos?
i.e. Cuantas personas sanas son identificadas como no enfermas.

Especificidad = $\frac{\text{verdaderos negativos}}{\text{elementos no relevantes}}$

Definir el éxito de una consulta

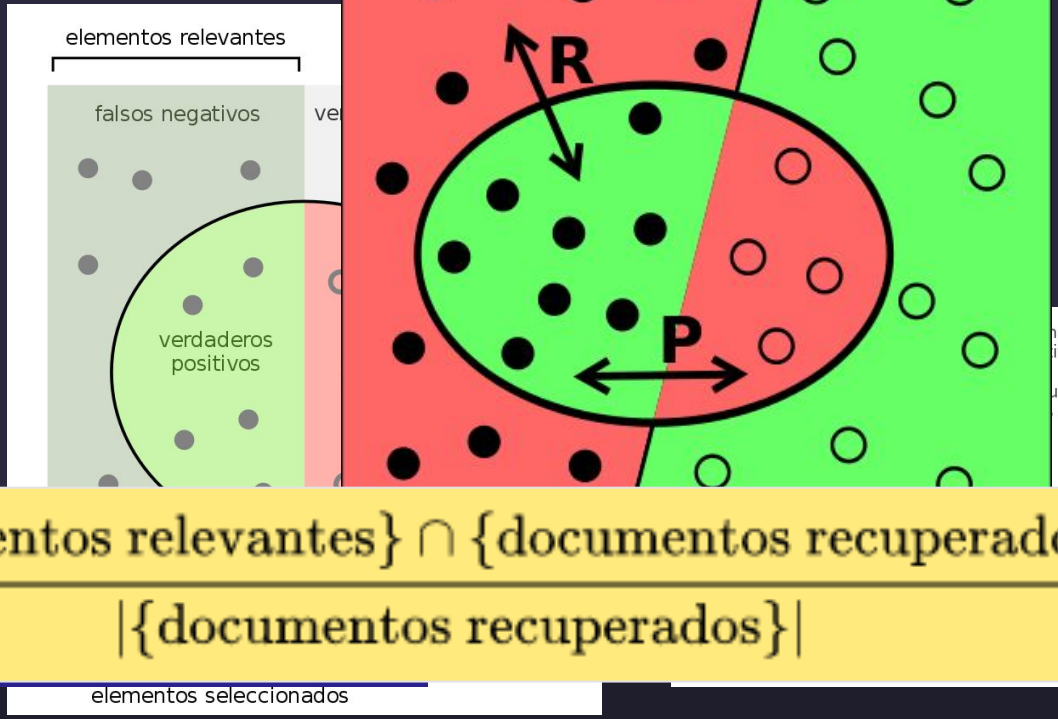
No es binario, sino b



Precisión y Exhaustividad (cobertura o *recall*)

Definir el éxito de una consulta

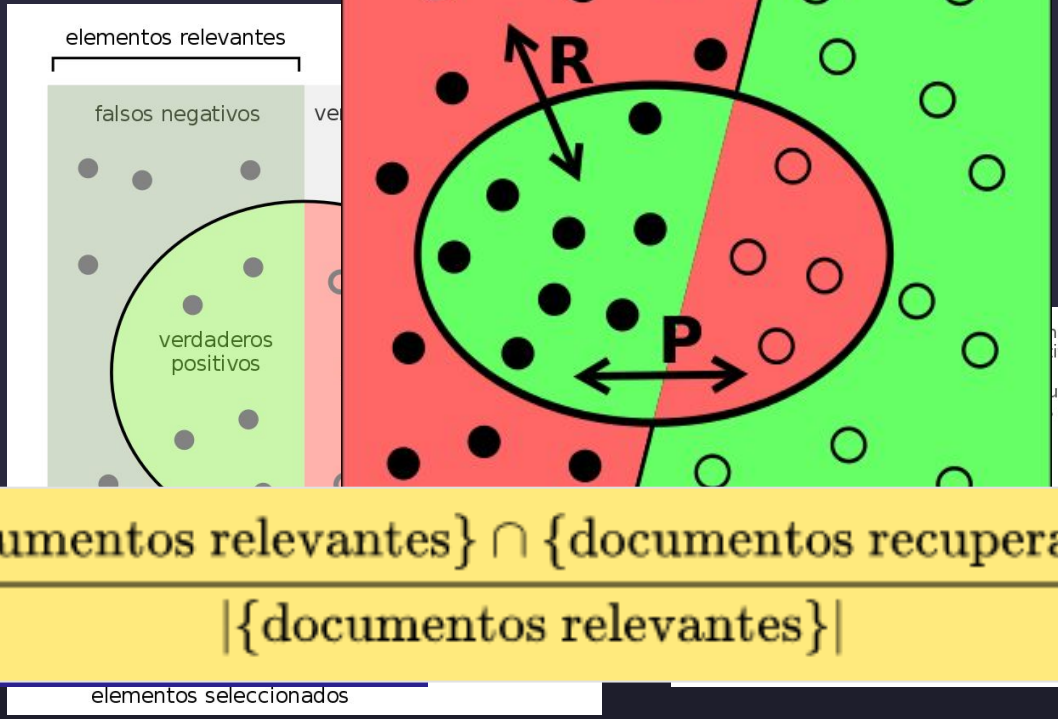
No es binario, sino b



$$\text{Precisión} = \frac{|\{\text{documentos relevantes}\} \cap \{\text{documentos recuperados}\}|}{|\{\text{documentos recuperados}\}|}$$

Definir el éxito de una consulta

No es binario, sino b



$$\text{Exhaustividad} = \frac{|\{\text{documentos relevantes}\} \cap \{\text{documentos recuperados}\}|}{|\{\text{documentos relevantes}\}|}$$

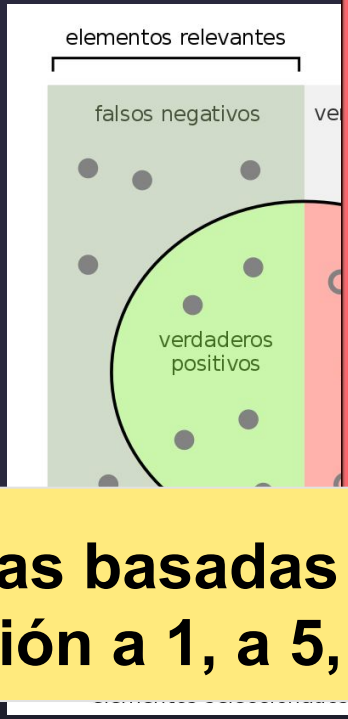
Definir el éxito de una consulta

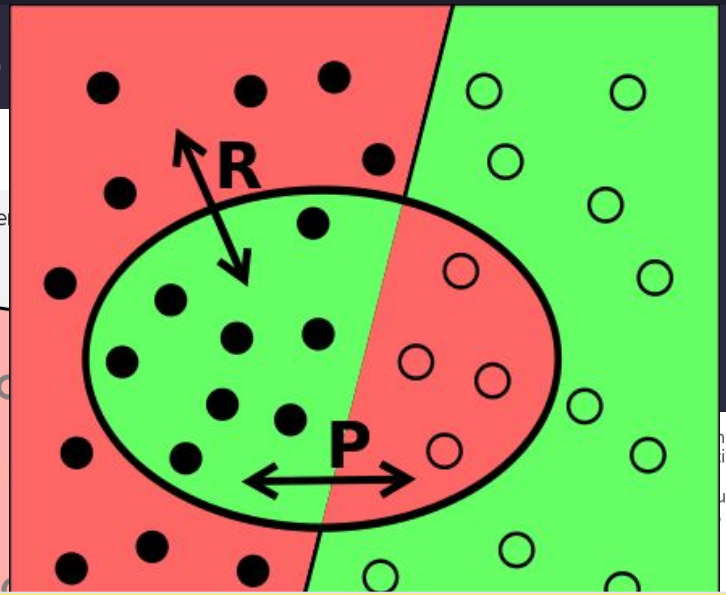
No es binario, sino b

$$F_1 = 2 \cdot \frac{\text{Precisión} \cdot \text{Exhaustividad}}{\text{Precisión} + \text{Exhaustividad}}$$

Definir el éxito de una consulta

No es binario, sino b





Métricas basadas en ranking:
Precisión a 1, a 5, a 10…

Evaluación en desafíos públicos: TREC

Con evaluación de especialistas

Datasets de referencia

Evaluación pública

Discusión de objetivos y métricas

Financiados por DARPA y NIST

¿Cómo lo hacemos?

Modelos basados en conjuntos: booleano

```
Tiene la palabra X
```

Modelos basados en conjuntos: booleano

Tiene la palabra X y Y pero no Z

La secuencia de palabras “X Y”

En el título, en el autor

Modelos algebraicos: Espacio vectorial

Cada documento es un vector,

cada palabra del vocabulario una dimensión

	Término 1	Término 2	Término 3
Documento 1	1	0	0
Documento 2	0	0	1
Documento 3	1	1	1
Documento 4	0	1	0

Modelos algebraicos: Espacio vectorial

Cada documento es un vector,

cada palabra del vocabulario una dimensión

Qué valores?

La palabra está en el documento

Cantidad de veces que ocurre la palabra

Probabilidad de ocurrencia de la palabra

Información mútua: tf*idf

(frecuencia de la palabra en el documento /
  frecuencia de la palabra en la colección)

Modelos algebraicos: Espacio vectorial

Cada documento es un vector,

cada palabra del vocabulario una dimensión

se calcula la **semejanza** entre documentos

(o entre documento y query)

como distancia en el espacio,

entre vectores

Modelos algebraicos: Espacio vectorial

Cada documento es

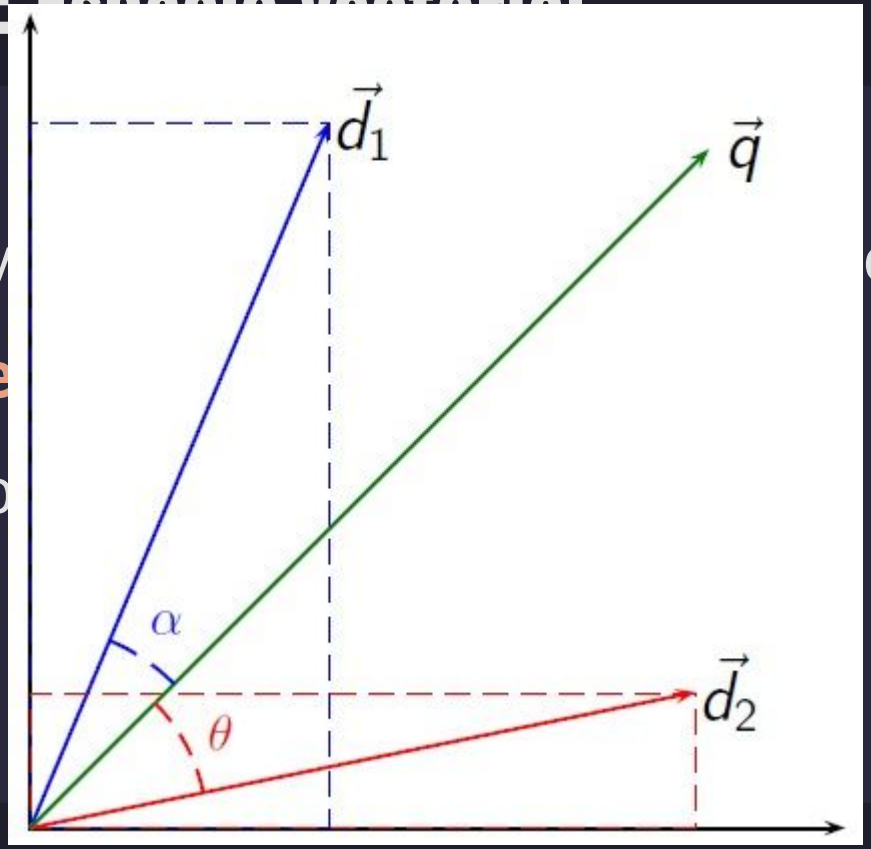
cada palabra del v ón

se calcula la **seme**

(o entre documento

como distancia en

entre vectores



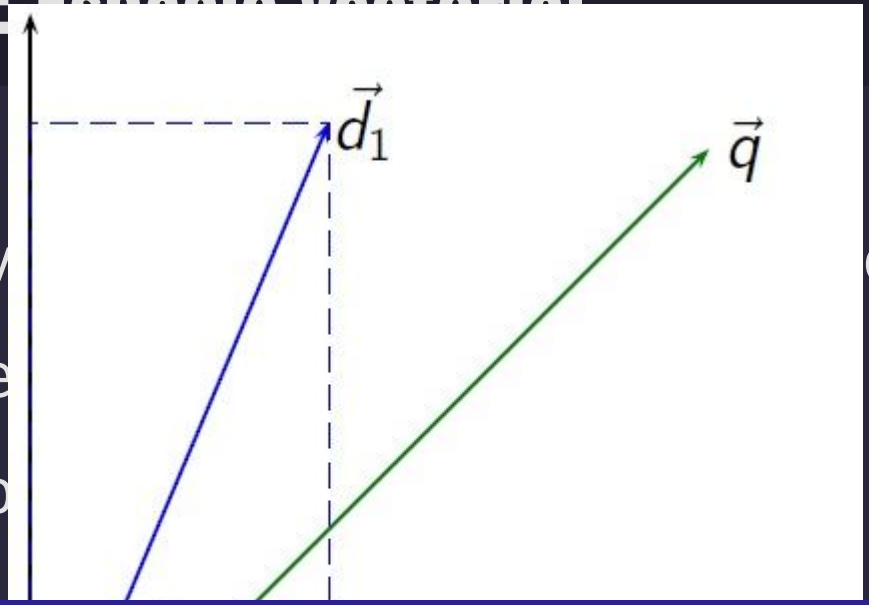
Modelos algebraicos: Espacio vectorial

Cada documento es

cada palabra del v ón

se calcula la seme

(o entre documento



Qué distancia?

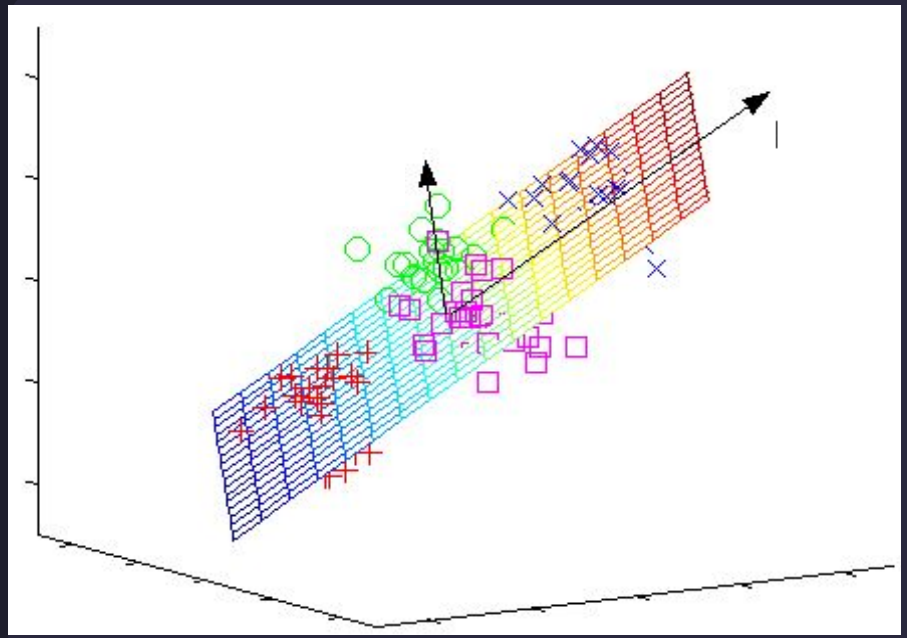
Coseno

Correlación (coseno normalizado por longitud)

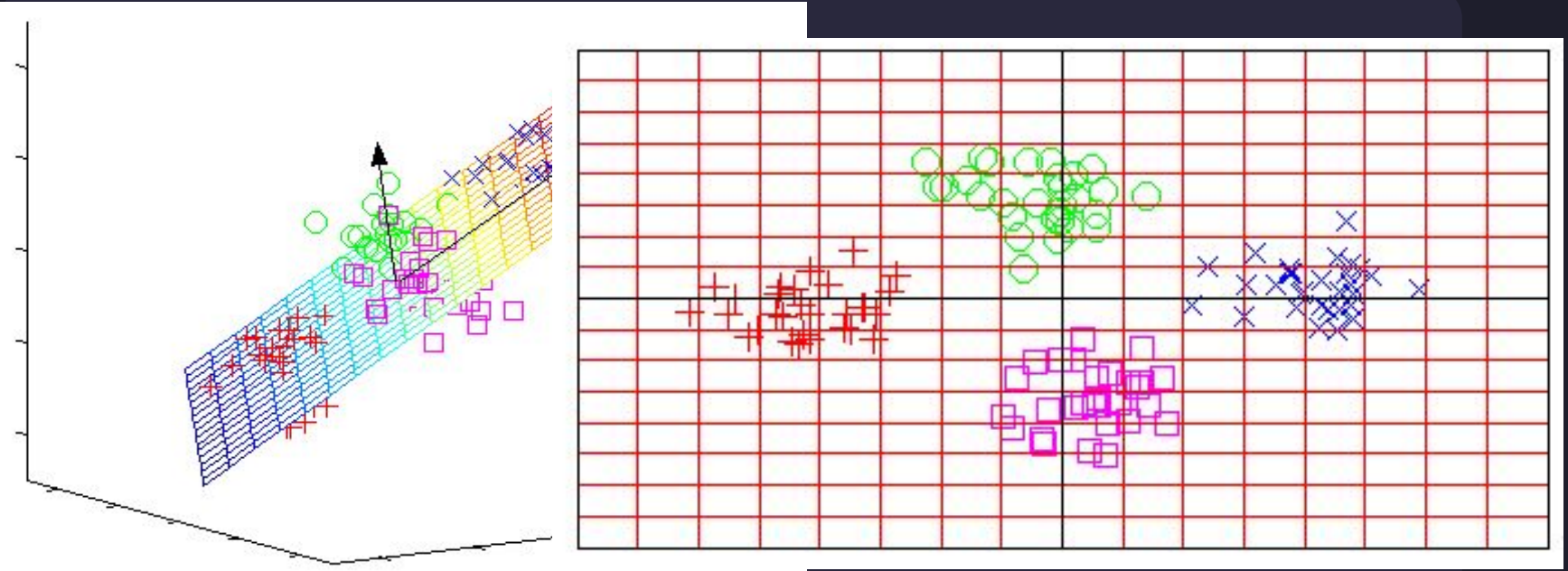
Semántica Latente

1. matriz documentos-términos
2. descomposición de valores singulares (PCA)
3. quedarse con las dimensiones de mayor variación

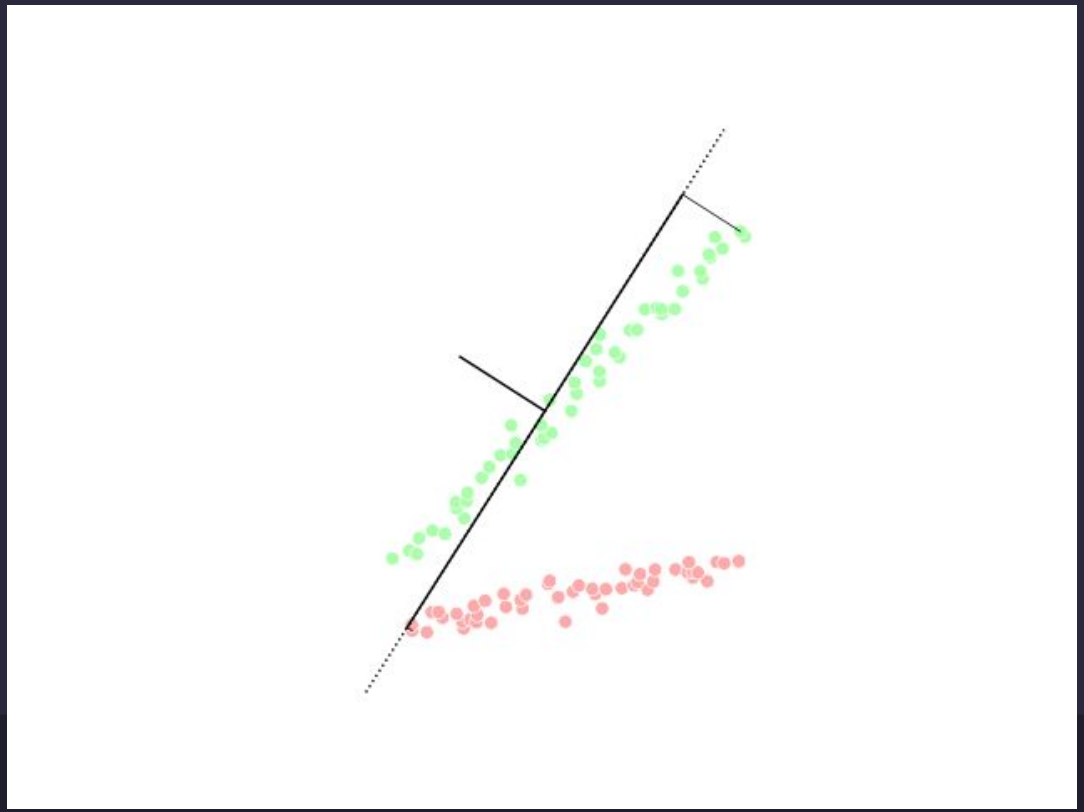
Semántica Latente: proyección



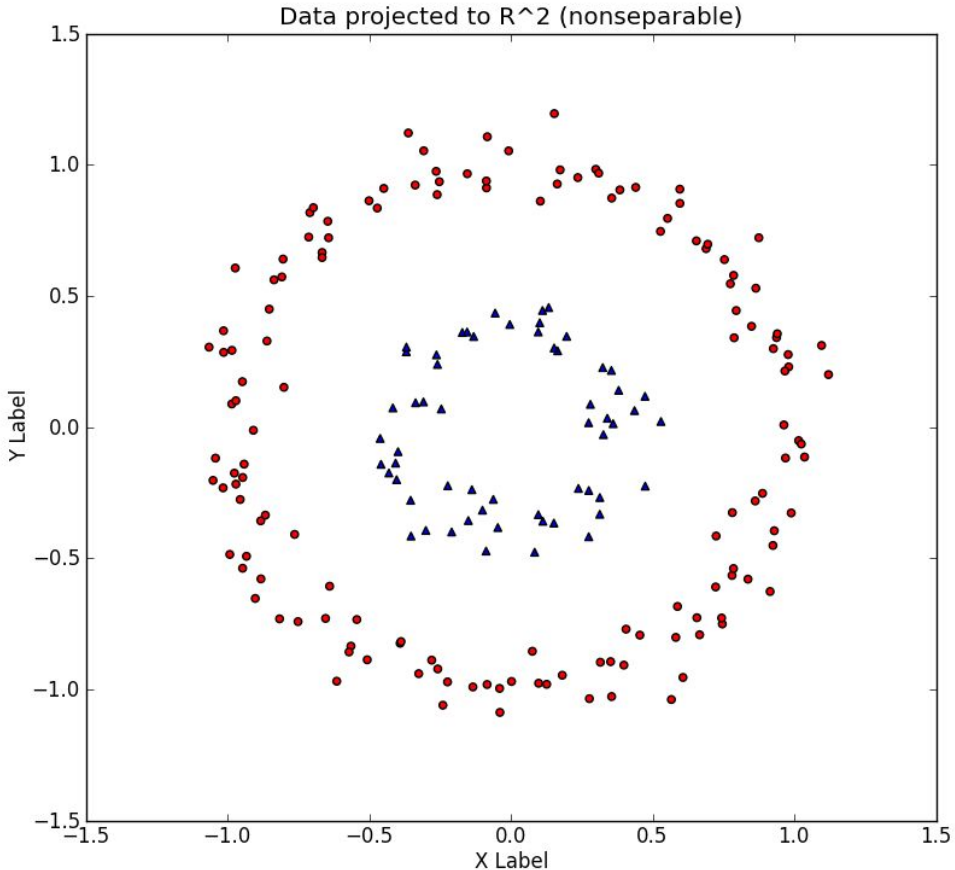
Semántica Latente: proyección

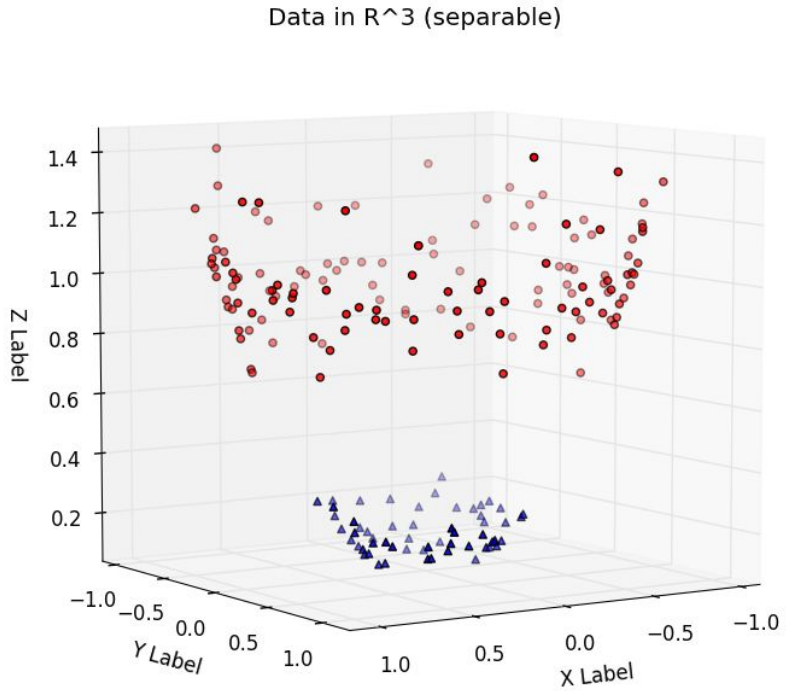


Semántica Latente: proyección



Semántica Latente: proyección





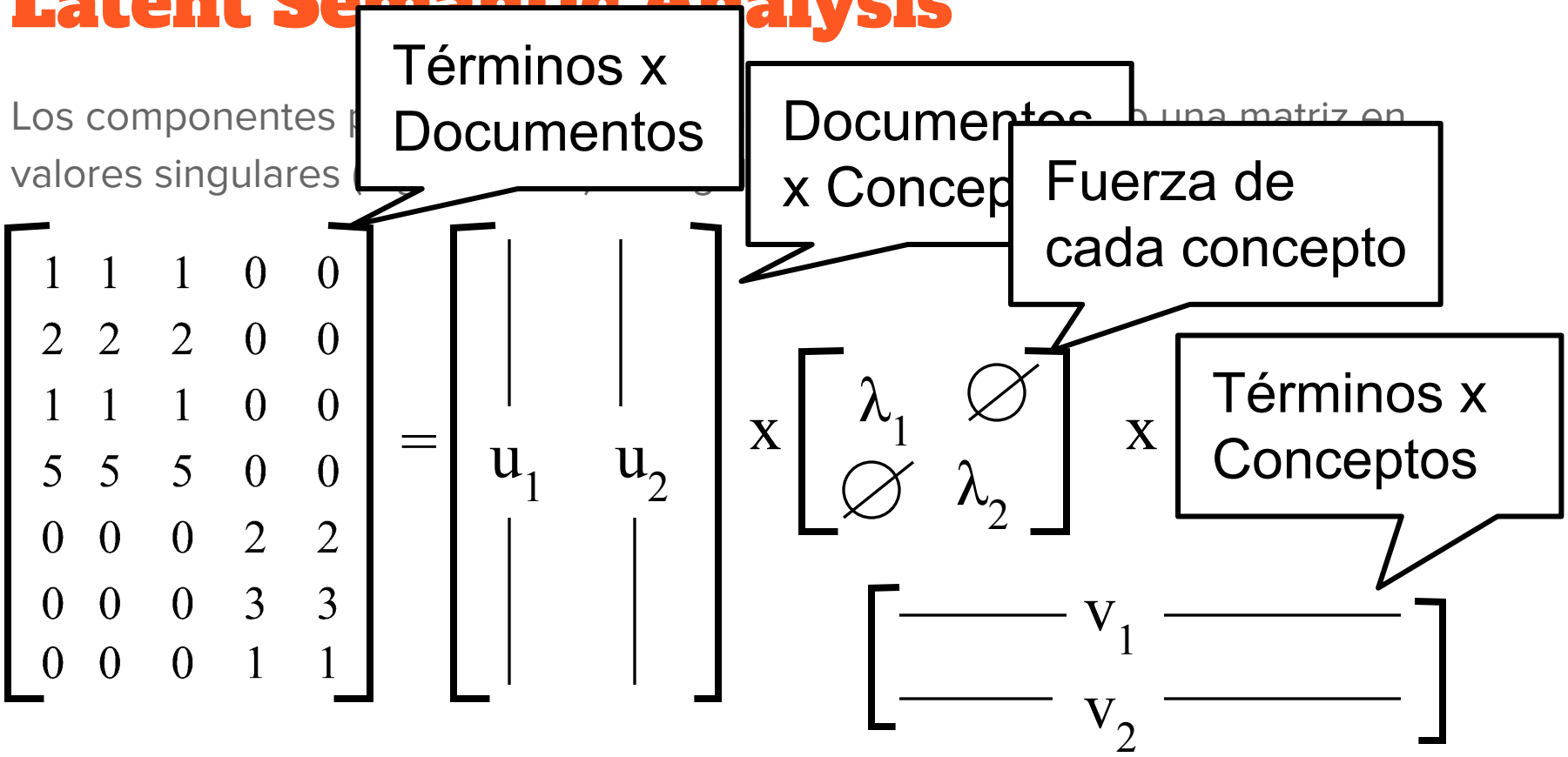
Descomposición en Valores Singulares

Los componentes principales se encuentran descomponiendo una matriz en valores singulares (eigenvalues) → singular value decomposition (SVD)

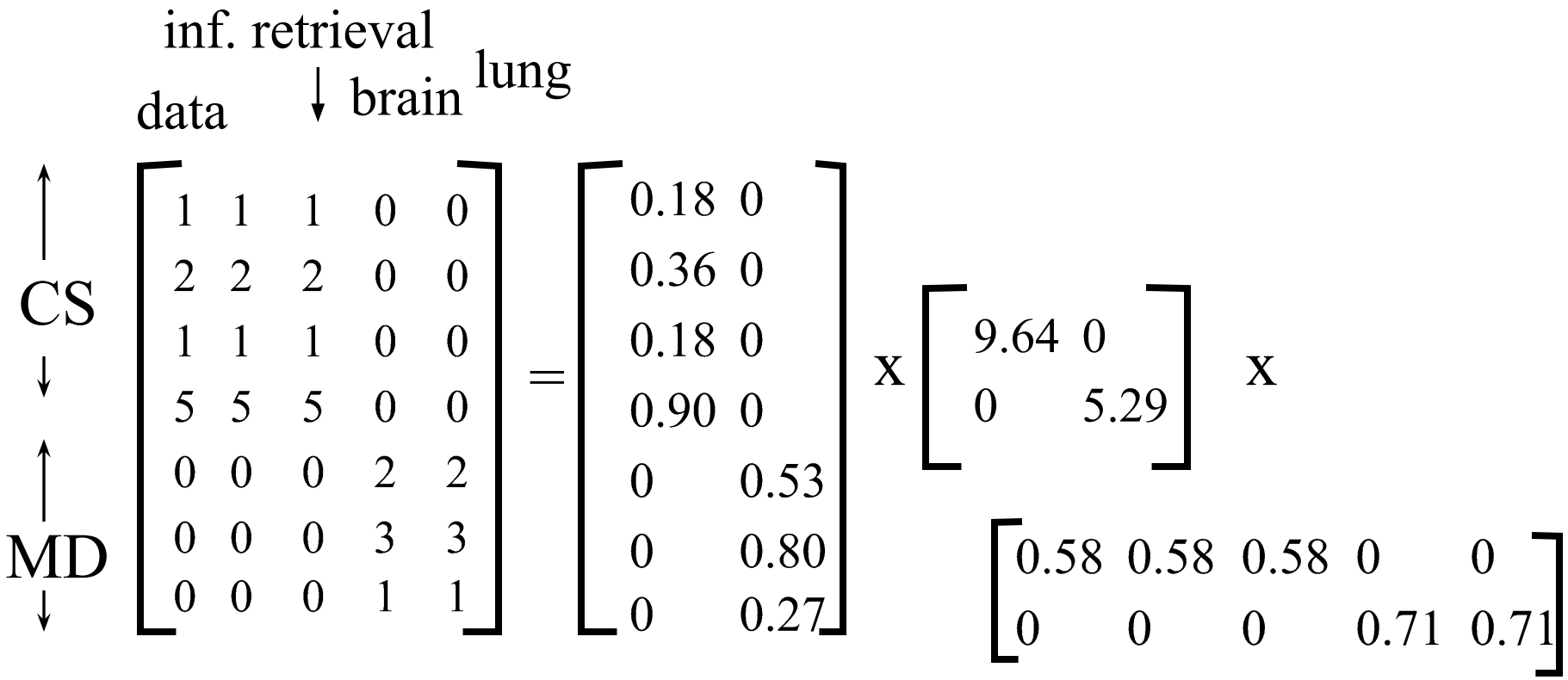
$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} | & | \\ u_1 & u_2 \\ | & | \end{bmatrix} \times \begin{bmatrix} \lambda_1 & \varnothing \\ \varnothing & \lambda_2 \end{bmatrix} \times \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \end{bmatrix}$$

Latent Semantic Analysis

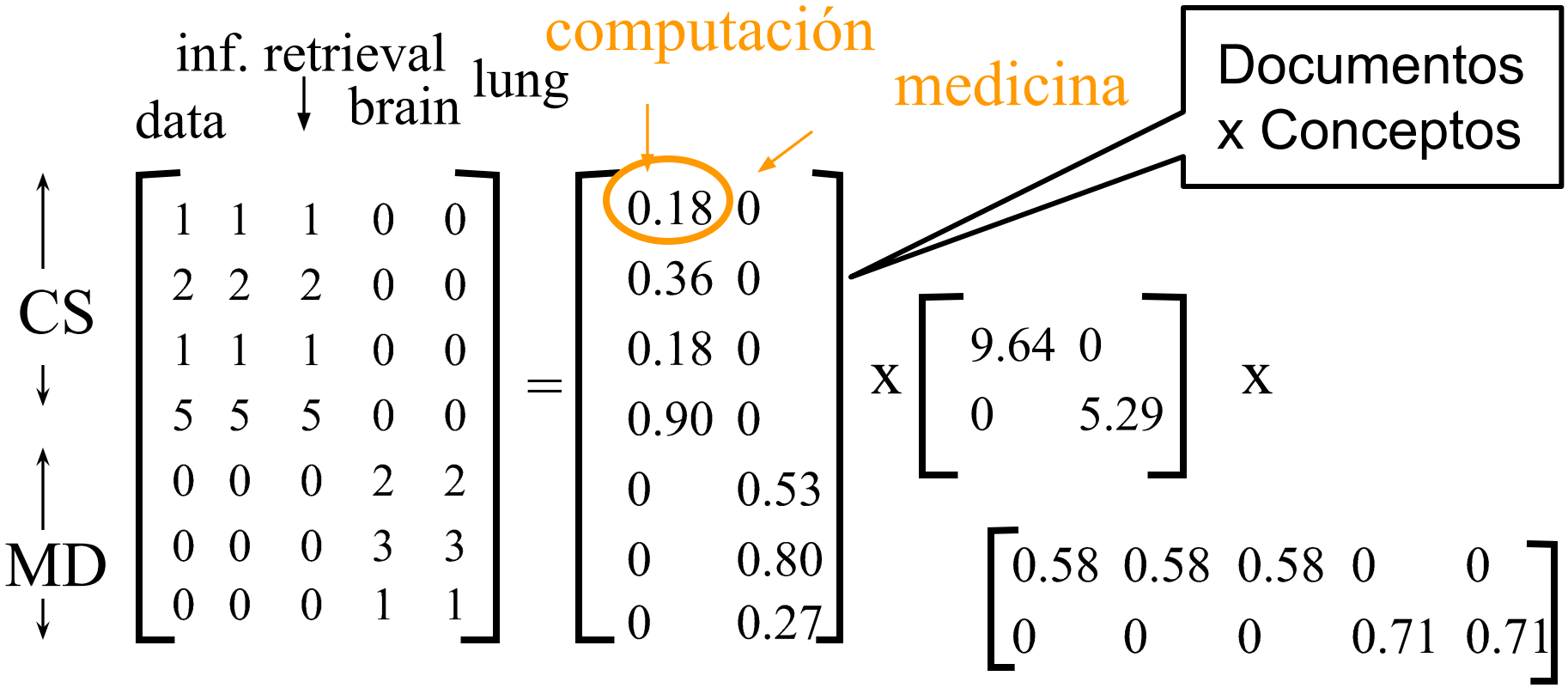
Los componentes p… una matriz en valores singulares (…



Latent Semantic Analysis

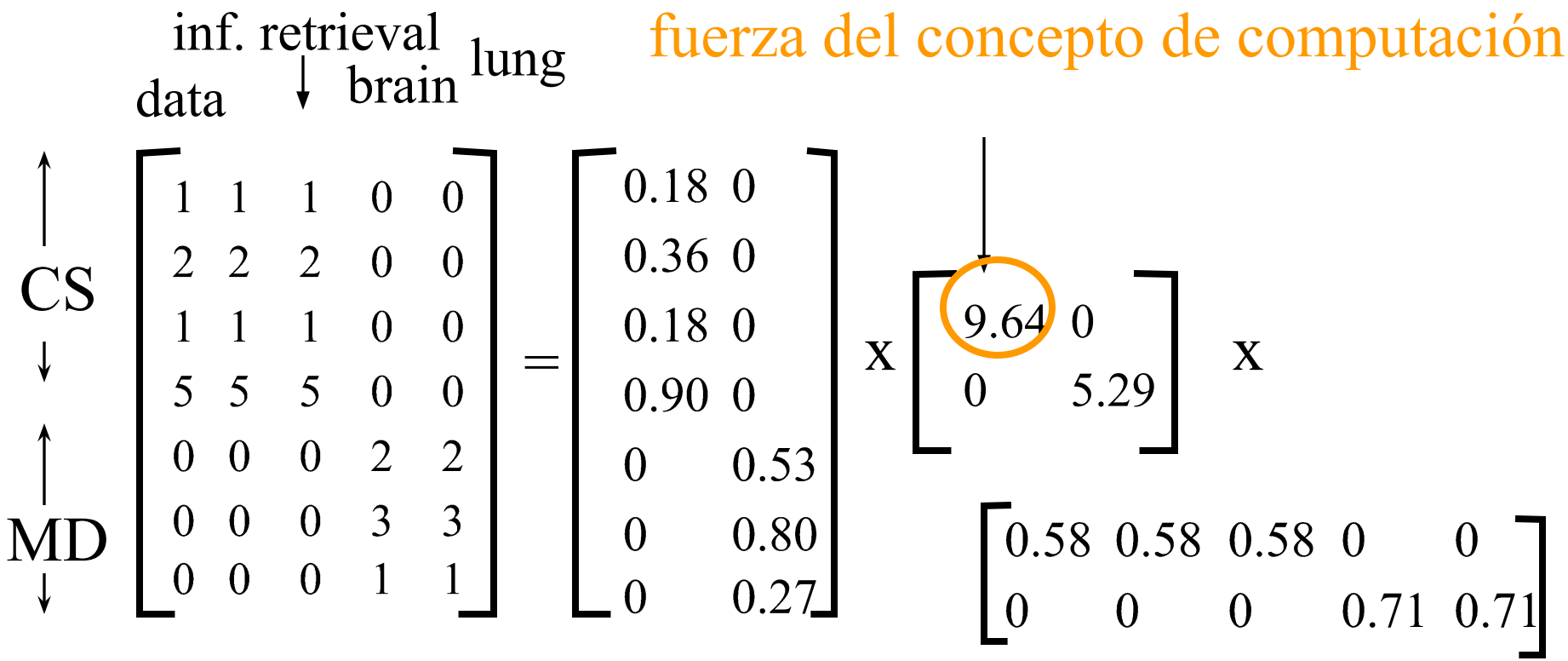


Latent Semantic Analysis

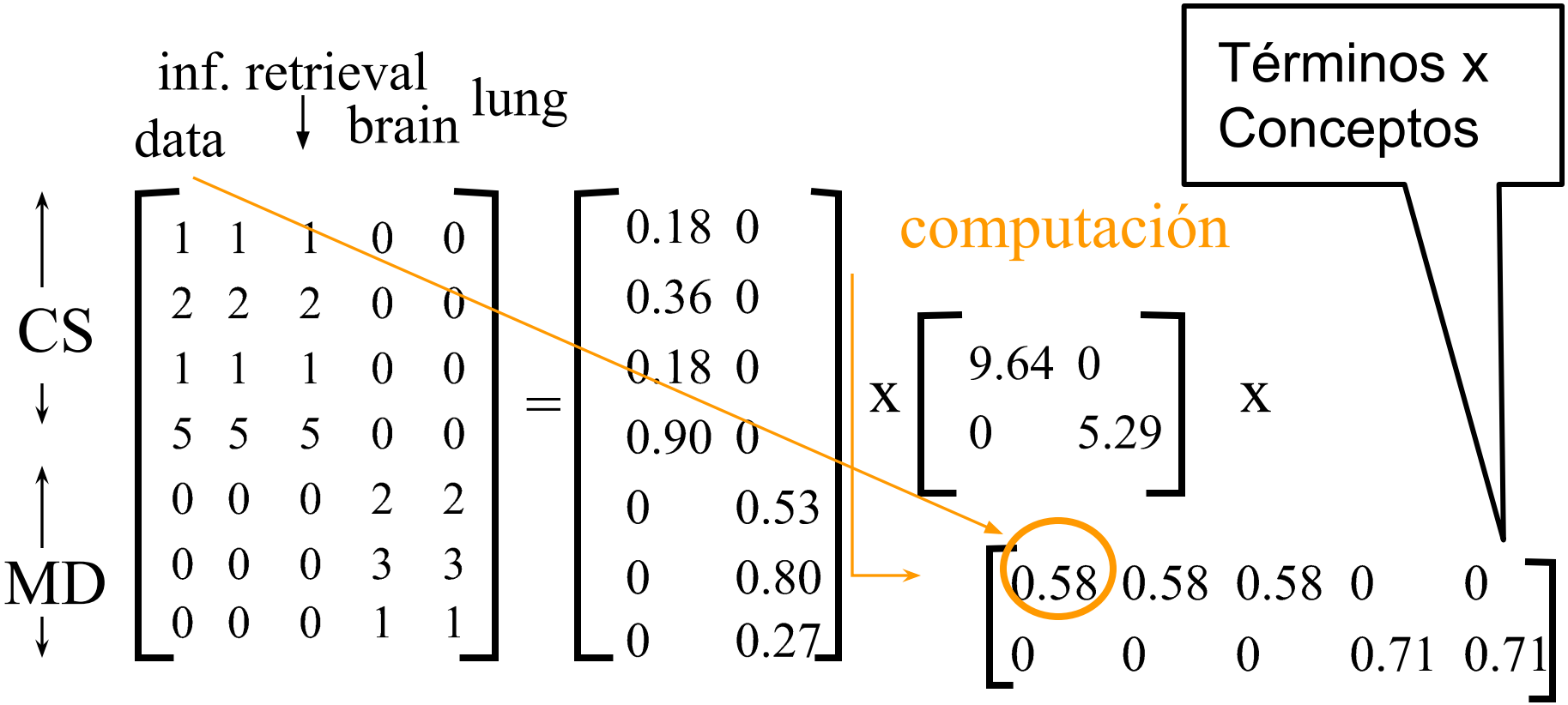


Latent Semantic Analysis

fuerza del concepto de computación



Latent Semantic Analysis



Latent Semantic Analysis: Reducción de dimensionalidad

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

Latent Semantic Analysis: Reducción de dimensionalidad

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \sim \begin{bmatrix} 0.18 \\ 0.36 \\ 0.18 \\ 0.90 \\ 0 \\ 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} 9.64 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \end{bmatrix}$$

Modelos algebraicos: Espacio vectorial con twist neuronal

Cada documento es un vector,

cada palabra del vocabulario una dimensión

se pasa por una red neuronal

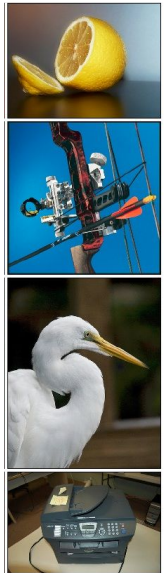
se obtienen nuevas dimensiones

se representan los documentos en esas dimensiones y… MEJORA

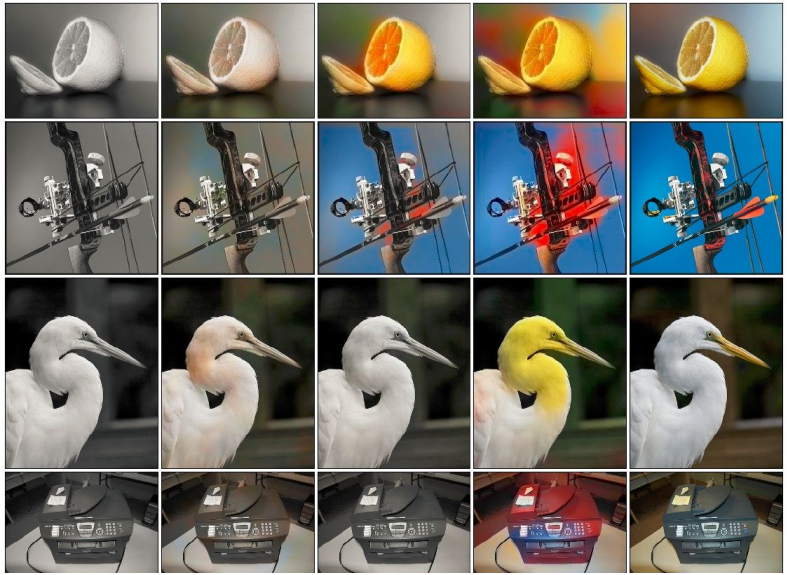
Embeddings neuronales

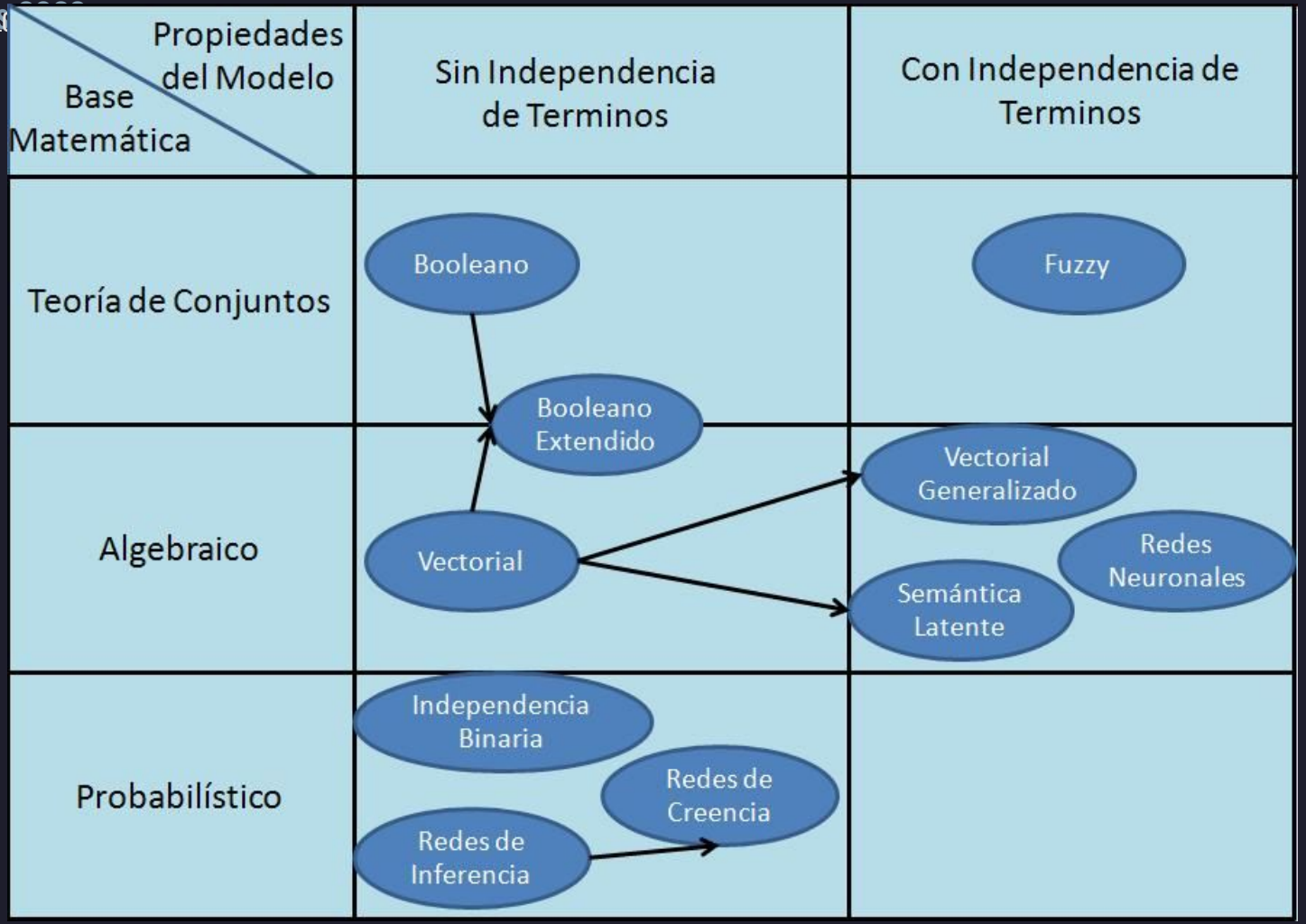
- Entrenar una red neuronal con una **tarea de pretexto** para la que tenemos muchos ejemplos naturalmente
  - Predecir una palabra dado su contexto, o un contexto dada una palabra
  - Reconstruir una imagen
- Eliminar la capa de predicción de la red
- La capa anterior a la de predicción es la nueva caracterización de los objetos
  - Menos características ➜ acercándonos a las causas latentes!
- Se usa la red para convertir los objetos del espacio original al espacio de embeddings
- Es relativamente barato de obtener
- Ahora podemos caracterizar datos supervisados con información poblacional de grandes cantidades de datos no supervisados

Tareas de pretexto



Tareas de pretexto




Propiedades del Modelo
Base Matemática
Sin Independencia de Terminos
Con Independencia de Terminos
Teoría de Conjuntos
Booleano
Fuzzy
Booleano Extendido
Algebraico
Vectorial Generalizado
Vectorial
Redes Neuronales
Semántica Latente
Probabilístico
Independencia Binaria
Redes de Creencia
Redes de Inferencia

¿Cómo es la infra de esto?

Lucene como indexador y recuperador

- librería de indexación y recuperación de texto
- basada en Documentos y Campos
- Apache, Java (originalmente)
- Independiente del formato del archivo, siempre que tenga texto

https://es.wikipedia.org/wiki/Lucene

Lucene como indexador y recuperador

- NO es un motor de búsqueda porque no crawlea la web ni parsea HTML → se puede integrar con Nutch para armar un buscador completo (search engine)

https://es.wikipedia.org/wiki/Nutch

¿Cómo se indexa?

1. Se obtienen las palabras relevantes para cada documento
   - todas
   - sólo las de contenido (eliminando “de”, “el”, “es”, etc.)
   - secuencias de dos, tres palabras

¿Cómo se indexa?

1. Se obtienen las palabras relevantes para cada documento
2. Se obtiene un índice invertido: para cada palabra, los documentos en los que aparece

¿Cómo se busca?

- Relevancia
- tf*idf (term frequency * inverse document frequency)
- semejanza (también recomendación): documentos que contienen las mismas palabras un número parecido de veces

¿Cómo se busca?

- semejanza semántica: documentos que contienen palabras semejantes un número parecido de veces
- sinónimos, probabilidades, distancia de edición

Esto puede ser muy grande

Wikimedia

CERN

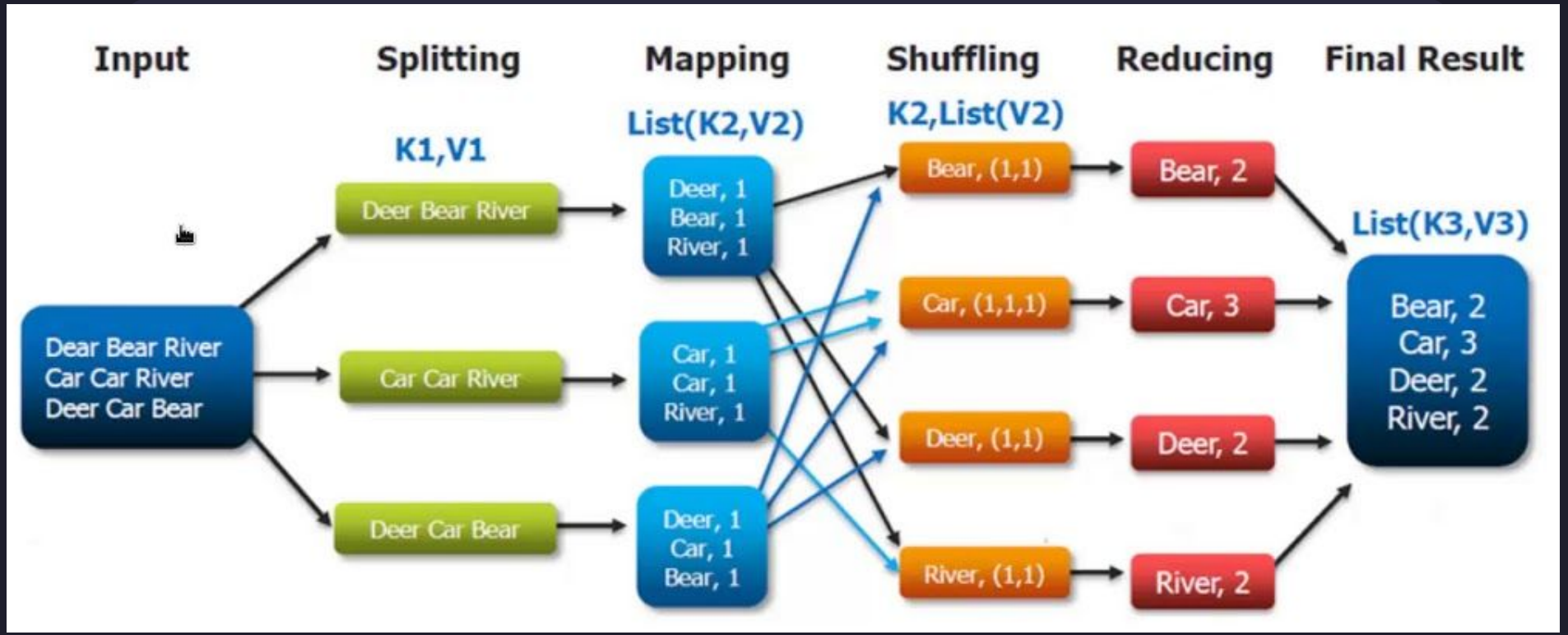
StackExchange

Github

Foursquare

La web!

Arquitectura distribuida



Arquitectura distribuida

Basada en map-reduce y sistemas de archivos distribuidos

- Hadoop
- Spark

Elasticsearch

Elasticsearch es un servidor de búsqueda basado en Lucene. Provee un motor de búsqueda de texto completo, distribuido, con una interfaz web RESTful y con documentos JSON. Elasticsearch está desarrollado en Java y está publicado como código abierto bajo las condiciones

Flujo Elastic

- datos sin procesar (logs, métricas de sistema, aplicaciones web, colecciones de documentos)
- ingesta de datos: parseo, normalización y enriquecimiento
- indexación
- consultas complejas (con un DSL)
- agregaciones
- visualizaciones (Kibana)

Ingesta con Logstash

Logstash es un pipeline de procesamiento de datos open source y del lado del servidor que te permite ingestar datos de múltiples fuentes simultáneamente y enriquecerlos y transformarlos antes de que se indexen en Elasticsearch.

Indexado con Lucene

- un índice de Elasticsearch es una colección de documentos JSON relacionados con un conjunto de claves (nombres de campos o propiedades) con sus valores correspondientes
- el índice invertido permite búsquedas de texto completo muy rápidas
- con la API de índice puedes agregar o actualizar un documento JSON en un índice específico.

Queries con Query DSL

- lenguaje simple, flexible y de gran alcance
- basado en JSON
- las consultas se componen de dos cláusulas:
- "Leaf Query Clauses" "match", "term" o "range", que devuelven un valor específico solicitado
- "Compound Query Clauses" combinación de Leafs para obtener información más compleja y detallada.

Ejemplos de queries

```
{
    "query":{
        "match" : {
            "color":"verde"
        }
    }
}
```

Ejemplos de queries

```
{
    "query":{
        "multi_match" : {
            "query": "montevideo",
            "fields": [ "ciudad", "departamento" ]
        }
    }
}
```

Visualización con Kibana

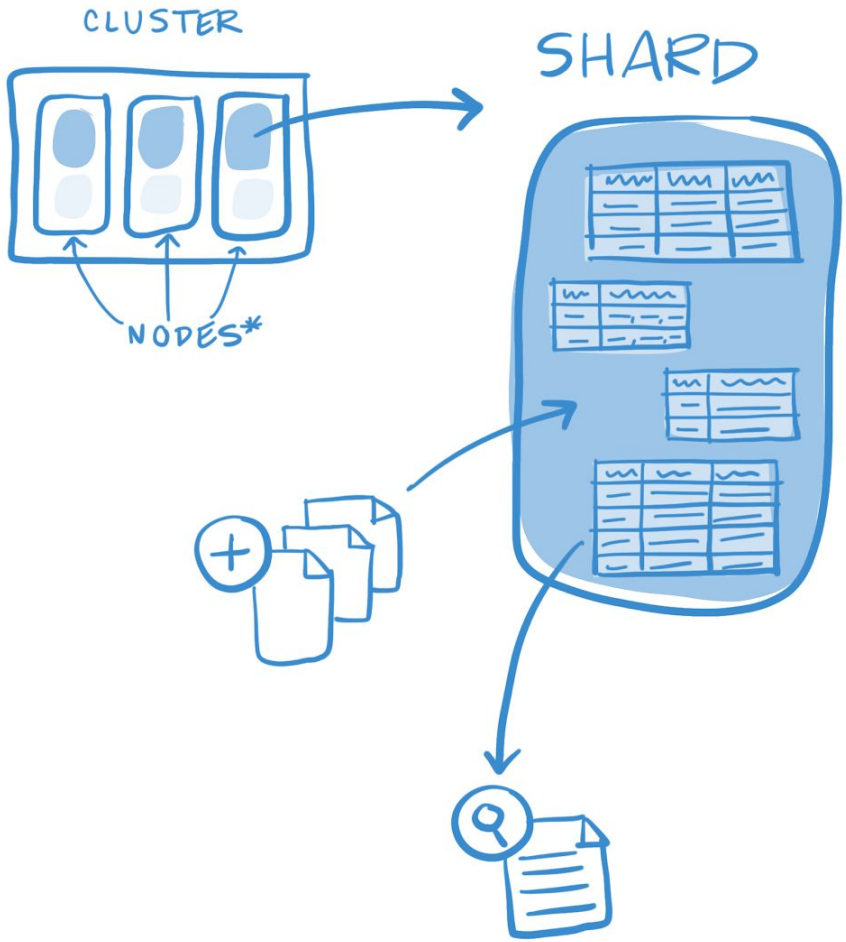
Kibana es una herramienta de visualización y gestión de datos para Elasticsearch que brinda histogramas en tiempo real, gráficos circulares y mapas.

Bonus

Muchas APIs Rest
https://www.elastic.co/guide/en/elasticsearch/reference/current/rest-apis.html
Analizadores de idioma (separación de palabras, listas de palabras vacías, identificación de números, etc.)
https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-lang-analyzer.html

Superbonus: Distribuido

los índices se pueden dividir en fragmentos (shards) y cada uno tener cero o más réplicas.
Cada nodo alberga uno o más fragmentos, actuando como un coordinador para delegar operaciones a los fragmentos correctos.
El rebalanceo y ruteo se realizan automáticamente.

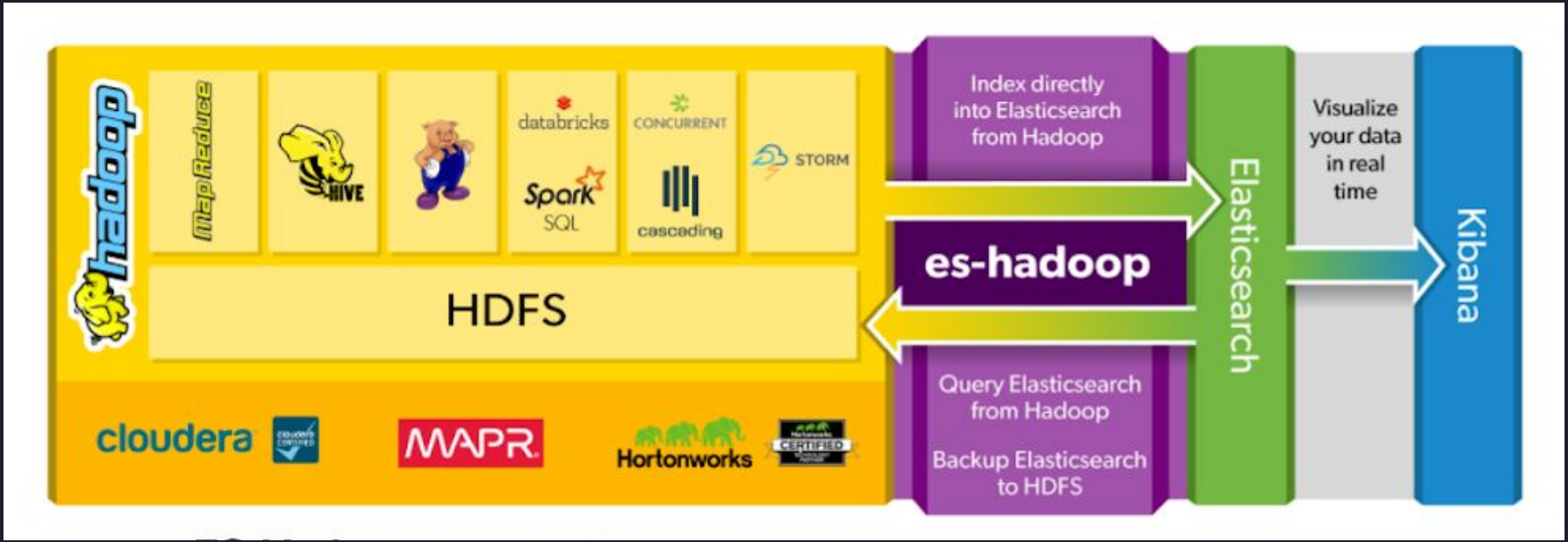


A SHARD IS
A LUCENE
INVERTED INDEX

- add documents
- search the index

*A **NODE** IS A
SERVER RUNNING
ELASTICSEARCH.
EVERY BONSAI
CLUSTER HAS
3 NODES.

Integrable con arquitecturas distribuidas



¿Otros stacks con las mismas funcionalidades?

Solr
Amazon Cloudsearch

Preguntas

¿Cómo indexar imágenes?
¿Cómo indexar sonido?
¿Cómo representamos palabras semejantes?

Principios ACID

ACID: Atomicity, Consistency, Isolation and Durability:

Atomicidad,

Consistencia,

aIslamiento y

Durabilidad.

Principios ACID

Atomicidad: transacciones completas

Consistencia: sólo se empieza lo que se puede acabar, si no se acaba, se revierte

aIslamiento: una operación no afecta otras

Durabilidad: persistencia

/THANKS!

/DO YOU HAVE ANY QUESTIONS?

youremail@freepik.com
+91 620 421 838
yourwebsite.com