

DATOS MASIVOS I

UNIDAD III MEDIDAS DE SIMILITUD Y DISTANCIA

MEDIDAS DE SIMILITUD Y DISTANCIA

¿Cómo Encontramos Imágenes en Internet?

house



Page 2 of about 413,000,000 results (0.08 seconds)

Related searches: [house tv show](#) [greg house](#) [house clipart](#) [cartoon house](#) [house music](#)



Click the Small **House** In The
465 × 346 - 58k - jpg
[supercoloring.com](#)
[Find similar images](#)



The **house** ...
600 × 400 - 93k - jpg
[museumoffloridahistory.com](#)
[Find similar images](#)



This large **house** ...
500 × 375 - 43k - jpg
[glamro.gov.uk](#)
[Find similar images](#)



HouseplanGuys.com, The largest
500 × 300 - 35k - jpg
[houseplanguys.com](#)
[Find similar images](#)

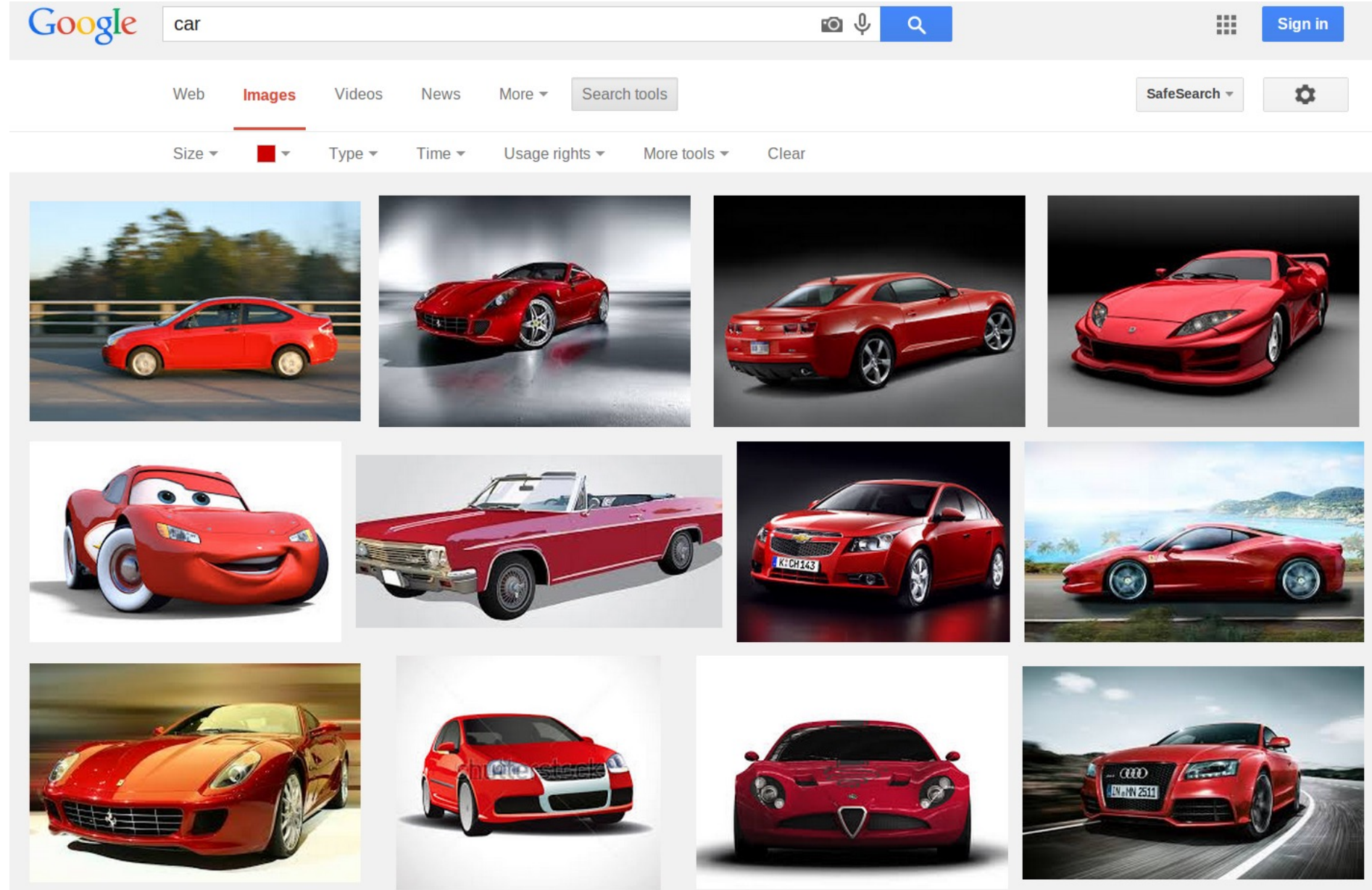


House picture by atkinson_crystal
800 × 600 - 299k - jpg
[s588.photobucket.com](#)
[Find similar images](#)

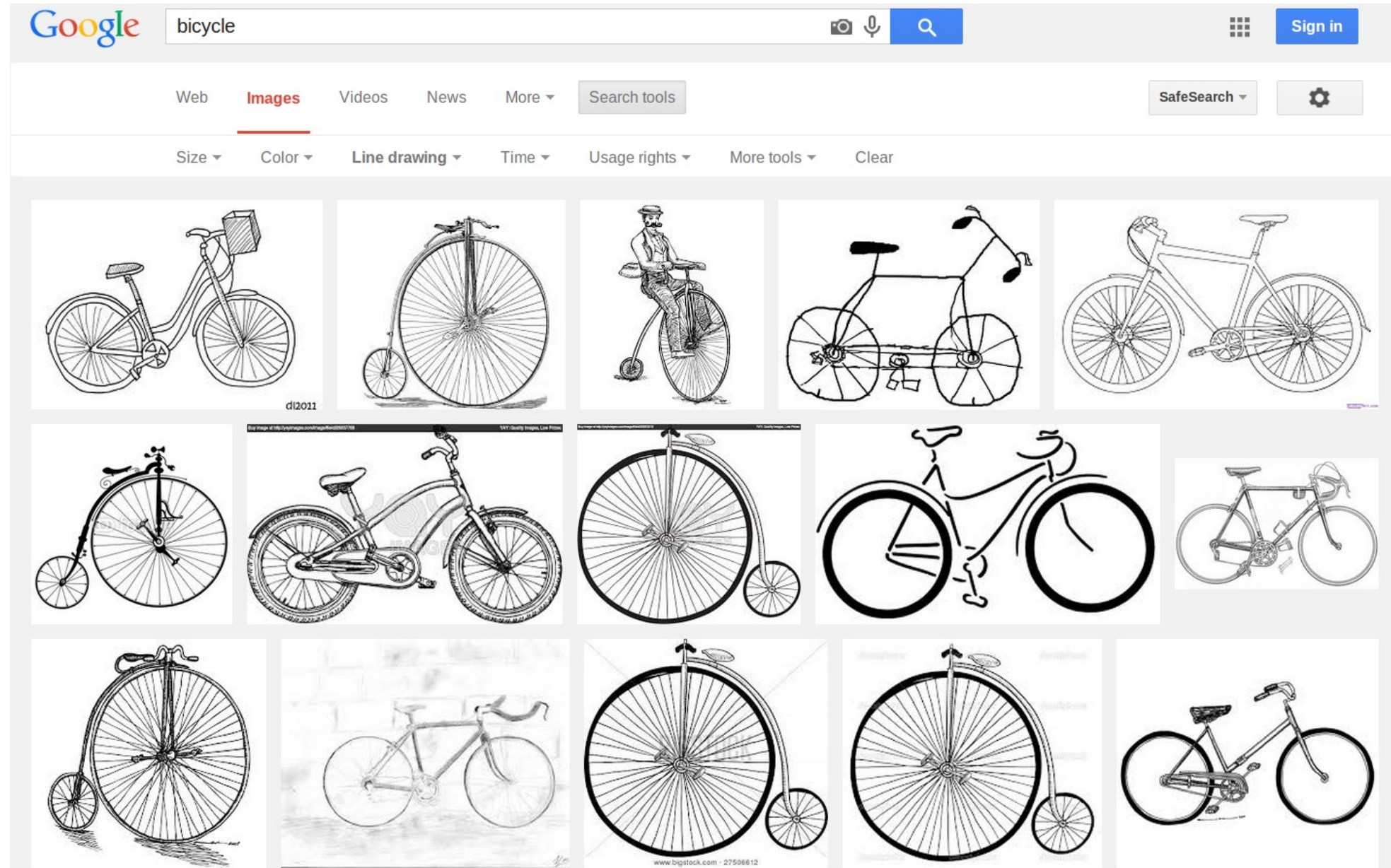


Barack & Michelle Obama P.
622 × 402 - 104k - jpg
[hiptics.com](#)
[Find similar images](#)





Búsqueda por Color




Búsqueda por Estilo

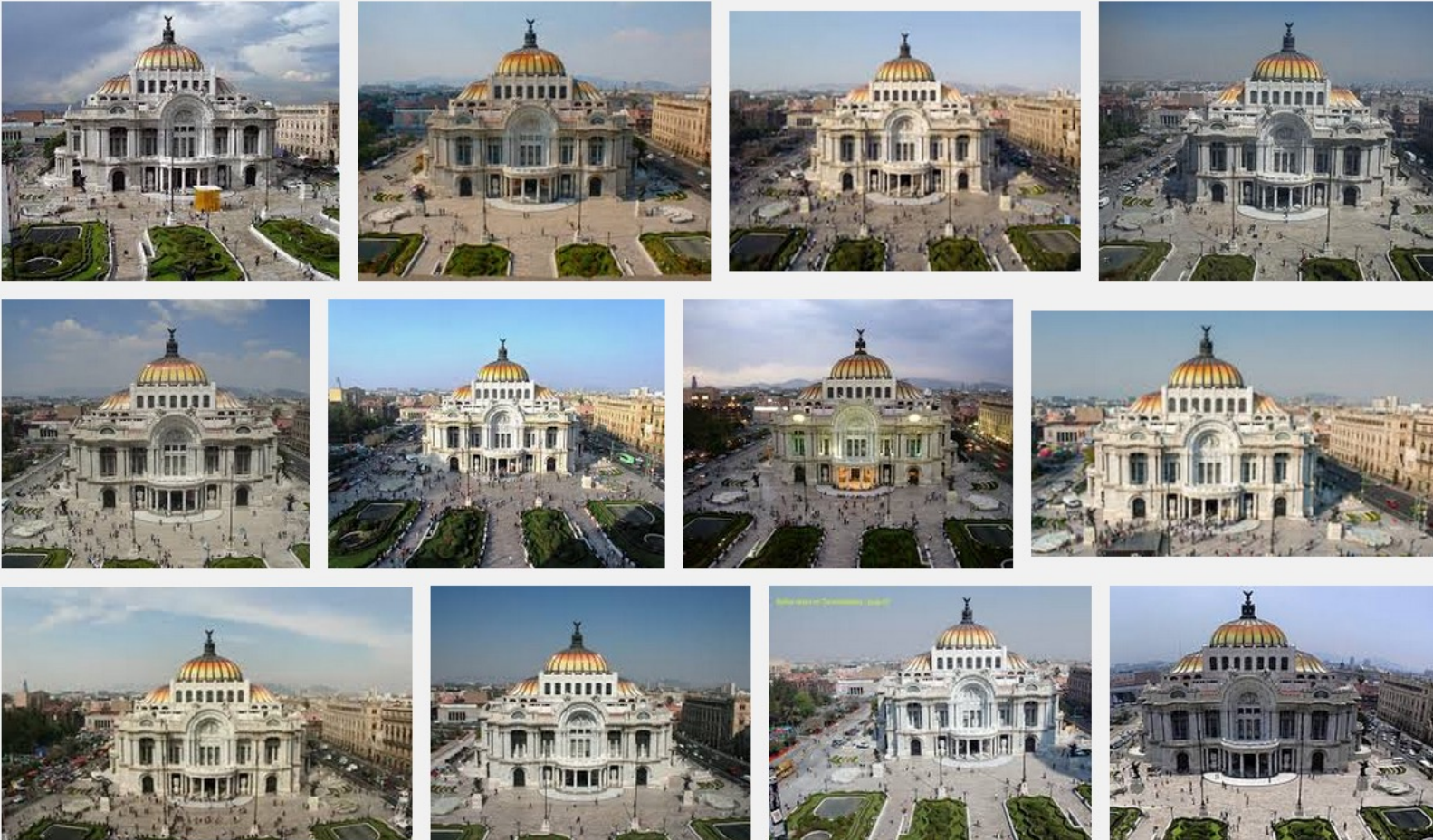


Búsqueda de Imágenes Visualmente Similares

Google     [Sign in](#)

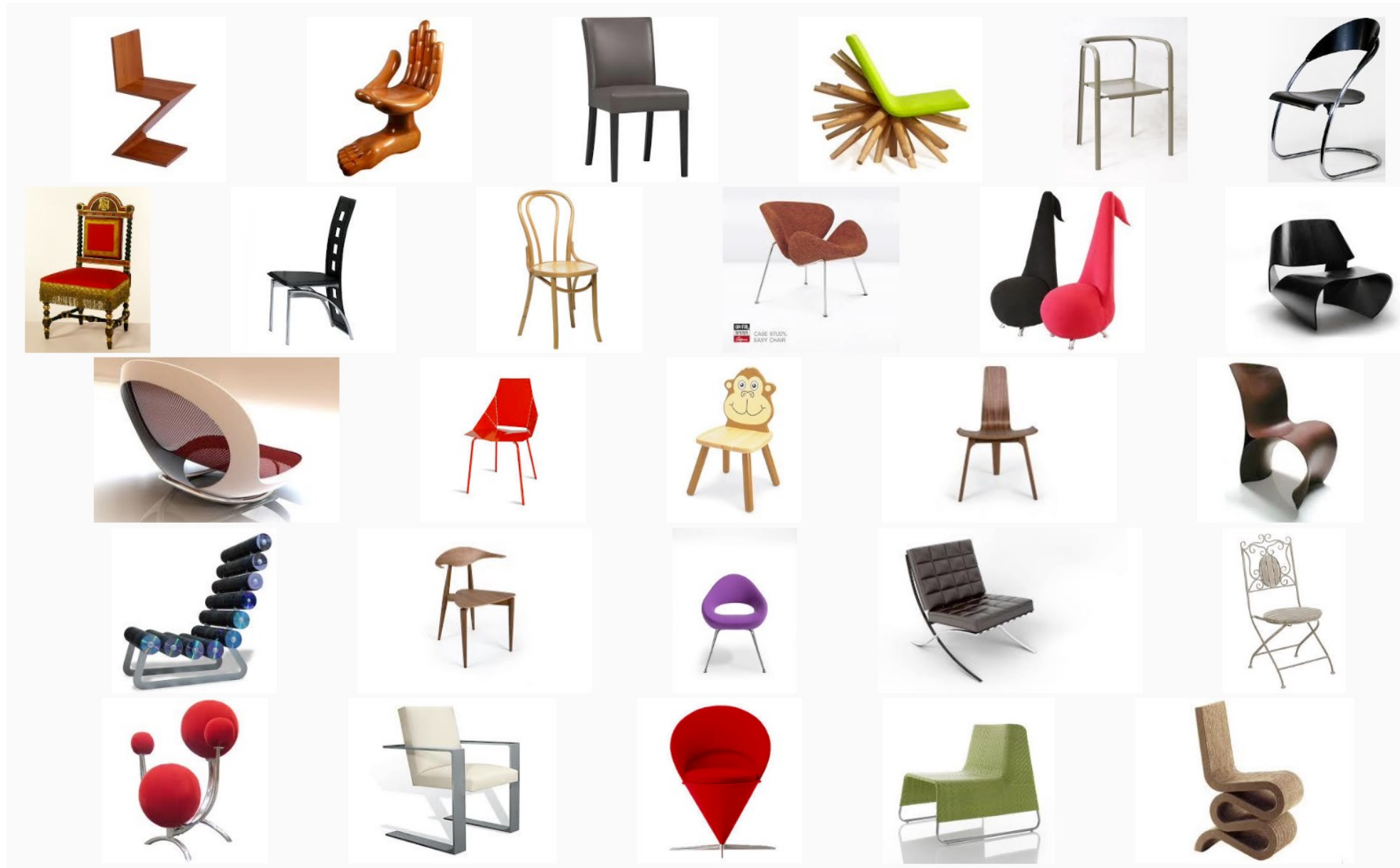
[Web](#) [Maps](#) [Images](#) [News](#) [Books](#) [More](#) 

[Size](#) [Color](#) [Type](#) [Time](#) [Visually similar](#) [Usage rights](#) [More tools](#) [Clear](#)



The image displays a Google search interface for visually similar images. The search bar is at the top, followed by navigation links (Web, Maps, Images, News, Books, More) and a 'Search tools' button. Below the navigation bar, there are filters for Size, Color, Type, Time, Visually similar, Usage rights, More tools, and Clear. The main content area shows a grid of 12 images, all of which are visually similar to the original image (a large, ornate building with a dome, likely the National Congress of Tucumán in Argentina). The images are arranged in three rows and four columns, showing different perspectives and lighting conditions of the same building.

Búsqueda de Imágenes de la Misma Categoría



¿Que hace que dos objetos sean iguales?

Similitud

¿Que hace que dos objetos sean iguales?



¿Estos textos son iguales?

Bosnia es la región geográfica mas grande con un clima continental moderado, marcado por veranos caluroso e inviernos fríos

Inslad es una región geográfica grande y tiene un clima continental moderado, caracterizado por veranos caluroso e inviernos fríos

¿Estos textos son iguales?

Bosnia es la región geográfica mas grande con un clima continental moderado, marcado por veranos caluroso e inviernos fríos

Inslad es una región geográfica grande y tiene un clima continental moderado, caracterizado por veranos caluroso e inviernos fríos

¿Podemos extraer algunas características de los dos objetos para determinar su grado de similitud?

Similitud

¿Estos textos son iguales?

Bosnia es la región geográfica mas grande con un clima continental moderado, marcado por veranos caluroso e inviernos fríos

Inslad es una región geográfica grande y tiene un clima continental moderado, caracterizado por veranos caluroso e inviernos fríos

¿Podemos extraer algunas características de los dos objetos para determinar su grado de similitud?

¿Cuáles métricas cuantitativas nos dicen qué tan similares son dos objetos?

Función de Similitud

- Cuantifica la similitud entre 2 objetos.
- Debería ser una métrica.

Función de Distancia

Debe ser una métrica, y cumplir con las propiedades:

- ✓ No negativo: $\text{distancia}(A, B) \geq 0$
- ✓ Identidad: $\text{distancia}(A, B) = 0$, si y solo si $A = B$
- ✓ Simetría: $\text{distancia}(A, B) = \text{distancia}(B, A)$
- ✓ Desigualdad del triángulo: $\text{distancia}(X, Y) \leq \text{distancia}(X, Z) + \text{distancia}(Z, Y)$

Distancias en la Precepción Humana

- No siempre se mantienen las propiedades de las distancias en la percepción humana. Por ejemplo, en la desigualdad del triángulo:

Distancias en la Precepción Humana

- No siempre se mantienen las propiedades de las distancias en la percepción humana. Por ejemplo, en la desigualdad del triángulo:

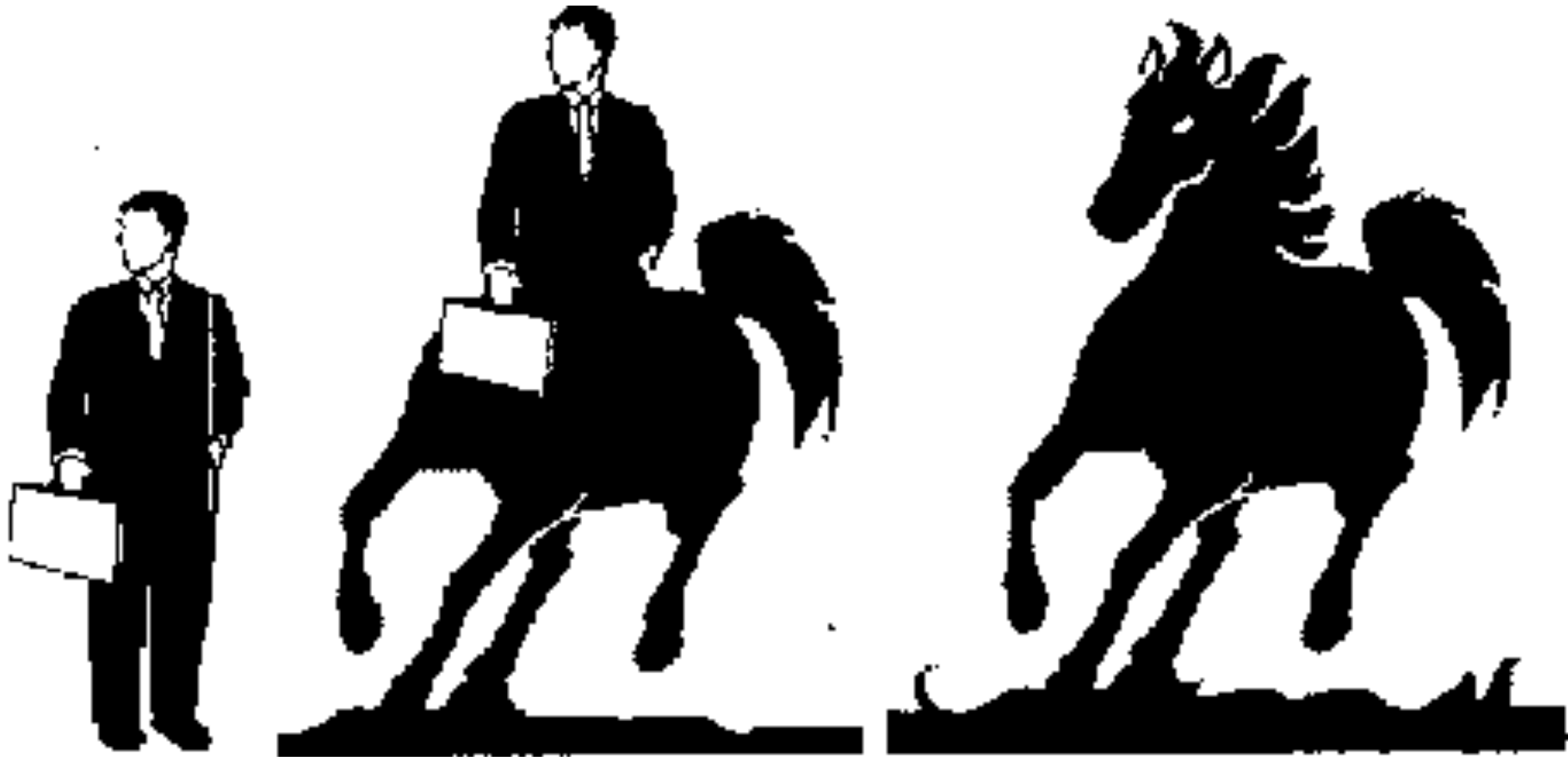
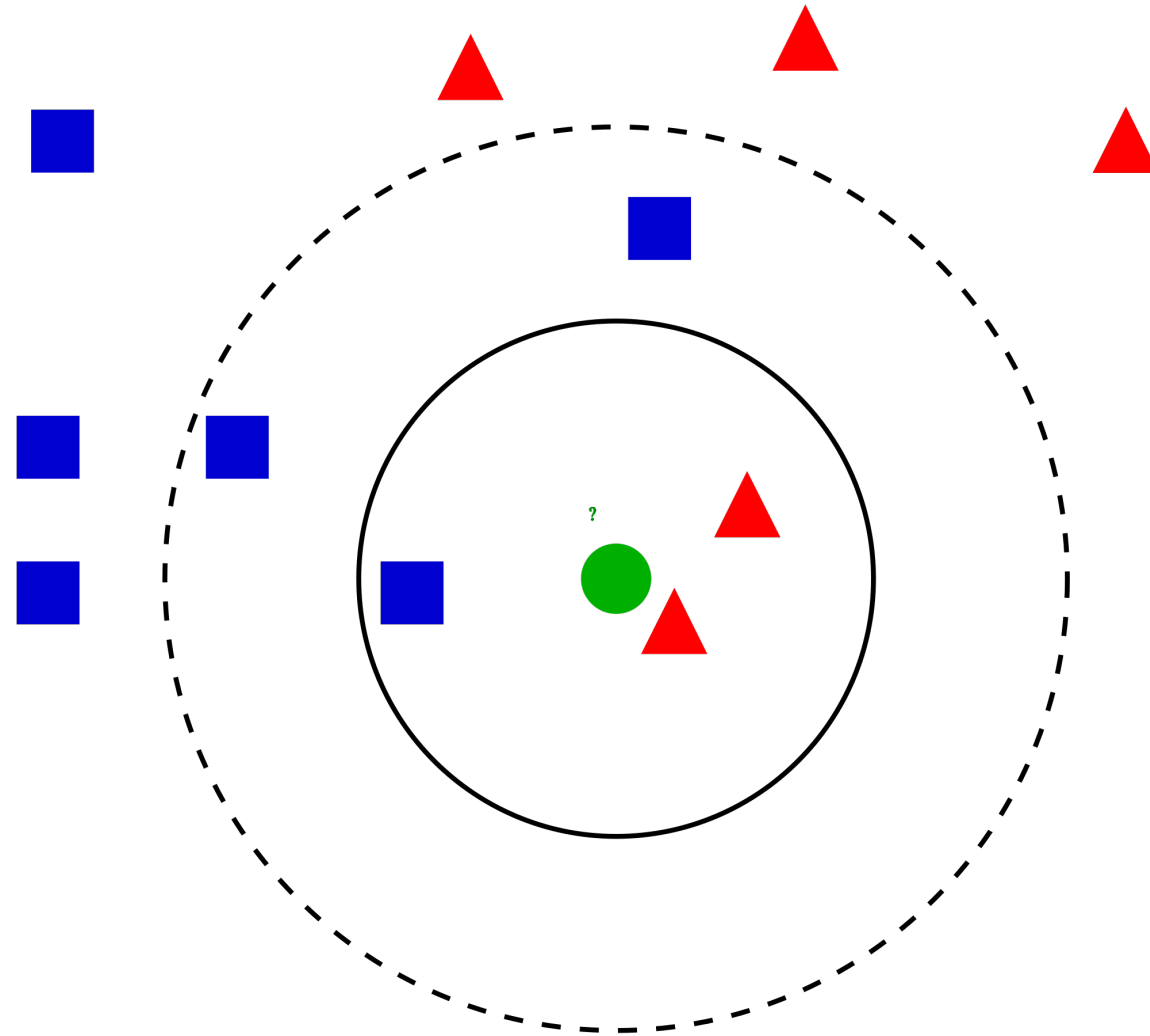


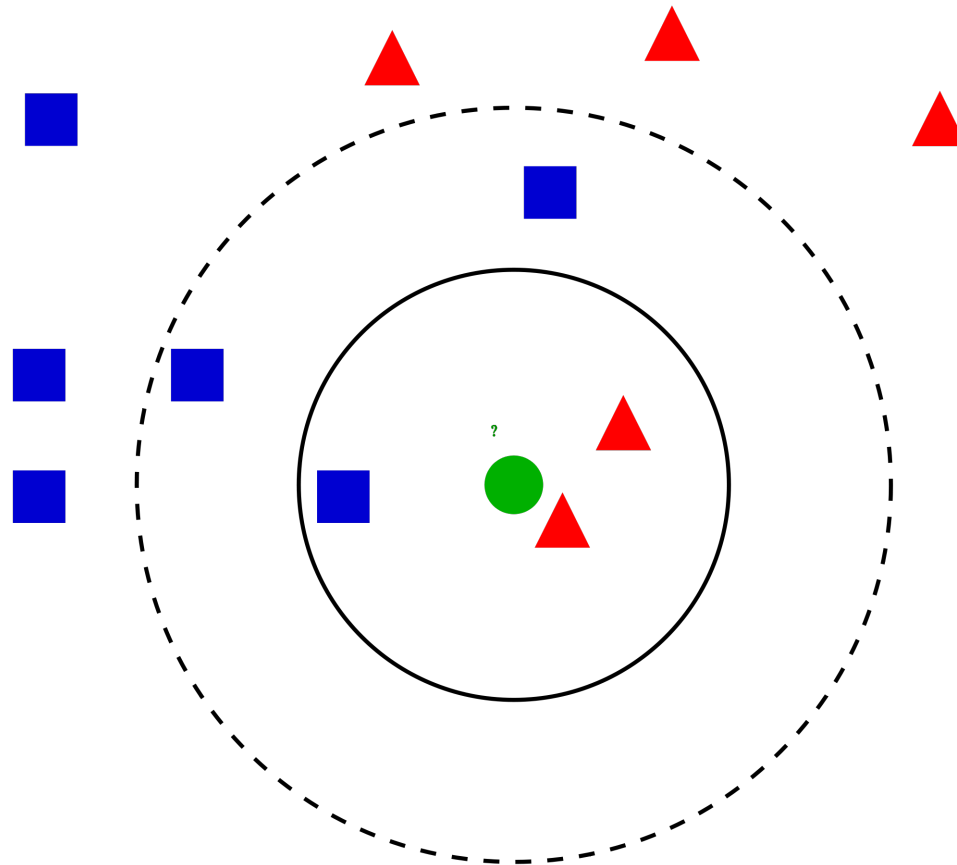
Imagen tomada de Veltkamp. Shape matching: similarity measures and algorithms, 2001.

Algoritmo K – *Nearest Neighbour* (Vecino Más Cercano)



Algoritmo K – *Nearest Neighbour* (Vecino Más Cercano)

El problema es encontrar el par de objetos $(x_1, x_2) \in X$ que son más similares o que son más cercanos bajo algún criterio de similitud o distancia $M(x_1, x_2)$.



El Problema del Vecino Más Cercano (*K – Nearest Neighbour*)

- Usando fuerza bruta requeriría comparar todos los pares posibles en X , lo cual es $\binom{n}{2} = \Theta(n^2)$.

El Problema del Vecino Más Cercano (*K – Nearest Neighbour*)

- Usando fuerza bruta requeriría comparar todos los pares posibles en X , lo cual es $\binom{n}{2} = \Theta(n^2)$.
- Se requiere tener todos los objetos en memoria para encontrarlo.

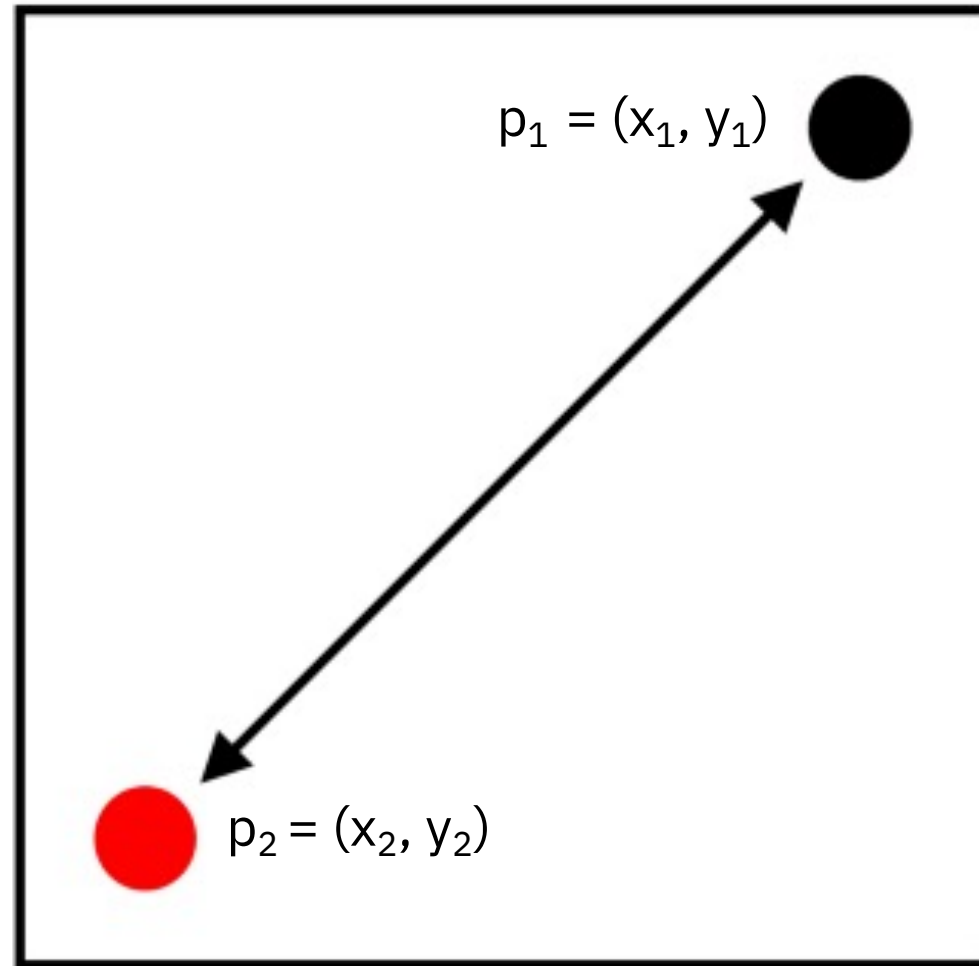
El Problema del Vecino Más Cercano (K – *Nearest Neighbour*)

- Tarea frecuente en análisis de datos (por ejemplo, en agrupamiento de clientes similares, búsqueda de documentos sobre el mismo tema, etc.).

El Problema del Vecino Más Cercano (K – *Nearest Neighbour*)

- Tarea frecuente en análisis de datos (por ejemplo, en agrupamiento de clientes similares, búsqueda de documentos sobre el mismo tema, etc.).
- Usado por algunos métodos no paramétricos de aprendizaje de máquinas (por ejemplo, el clasificador de *k-vecinos* más cercanos.).

Distancia Euclidiana

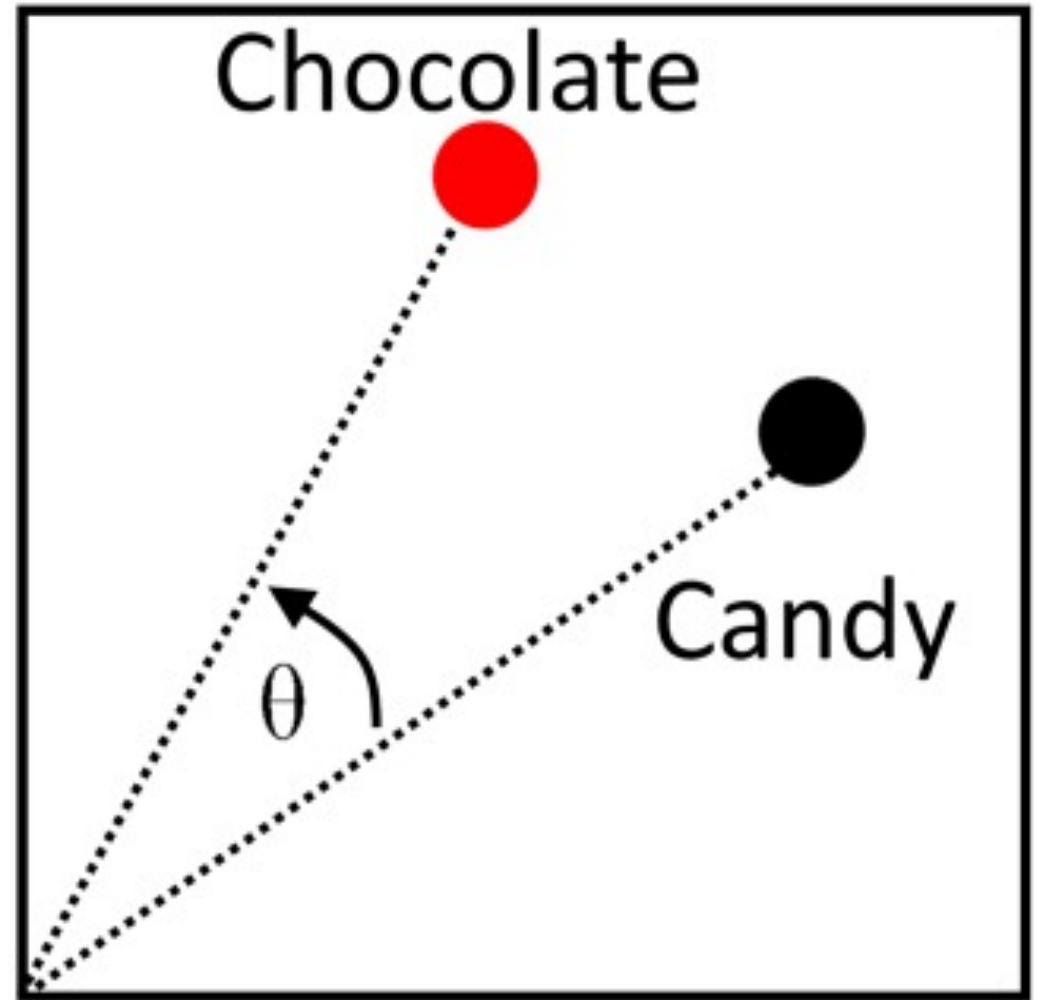


$$Distancia(x, y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Distancia Coseno

La similitud coseno compara la orientación de 2 vectores mediante el coseno del ángulo entre ellos.

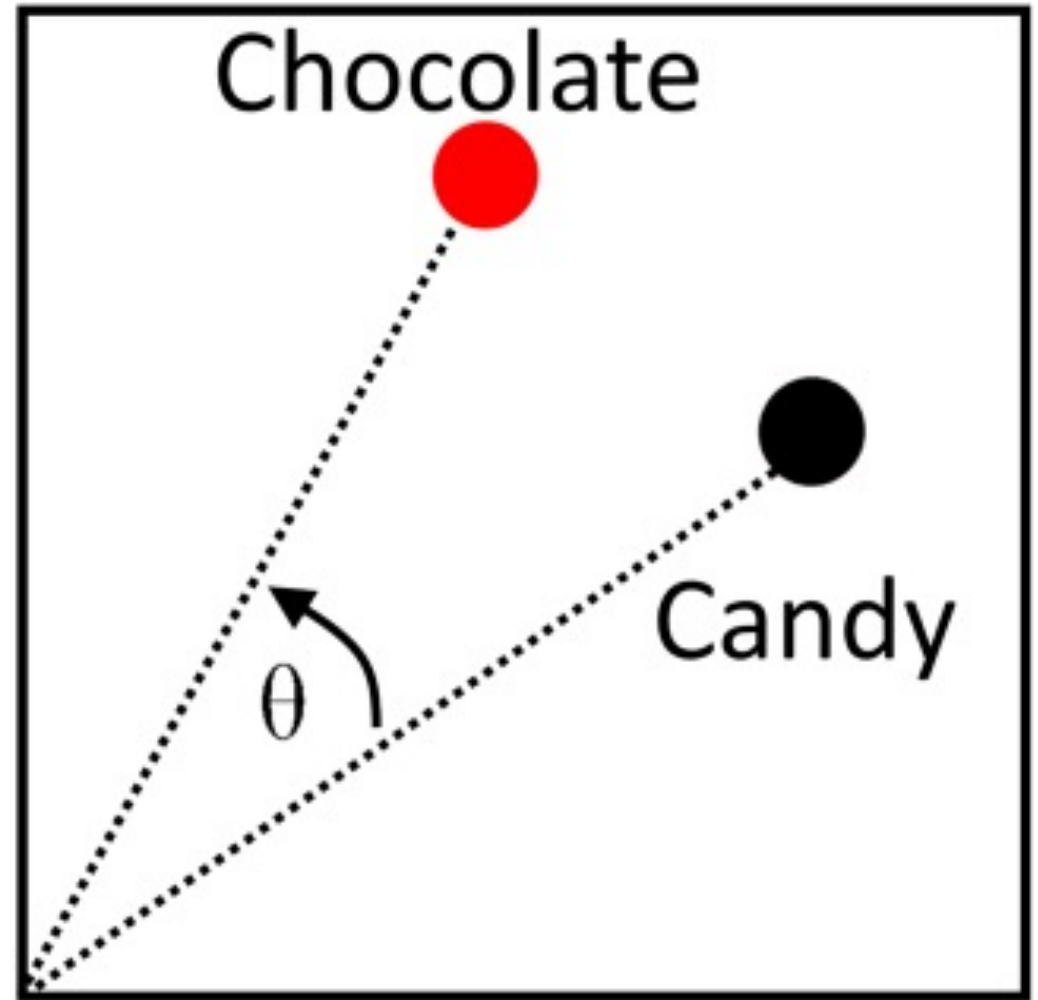
$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$



Distancia Coseno

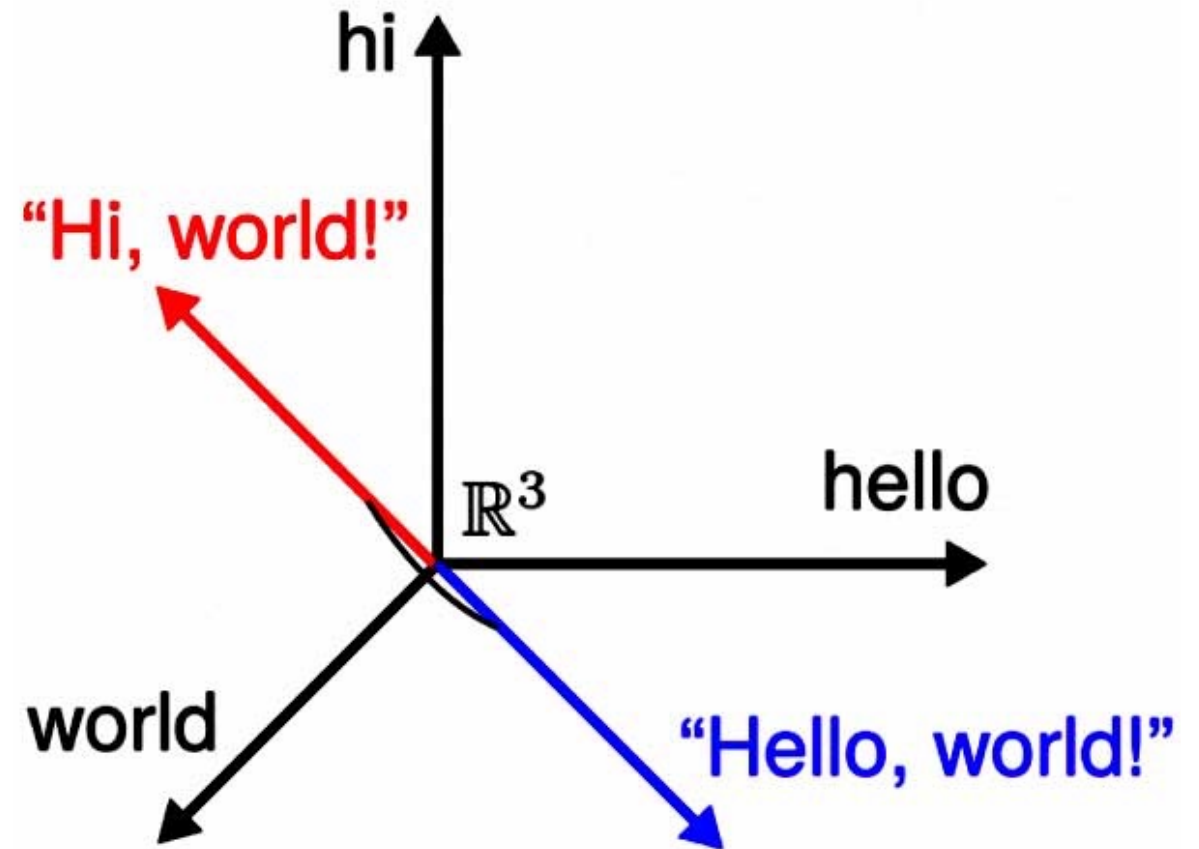
Dos vectores con la misma orientación tienen una similitud de 1.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$



Distancia Coseno

Comúnmente usada
para comparar
documentos de
texto.



Distancia de Hamming

- Para un tamaño fijo T , es el número de elementos distintos de 2 vectores o cadenas

Distancia de Hamming

2	5	9	1	6	7	4
2	0	9	1	5	7	4

2

B	O	C	I	N	A	S
C	O	M	P	R	A	R

5

1	0	1	1	1	1	0
1	1	0	1	0	1	0

3

Calcula la distancia de Hamming de los siguientes objetos:

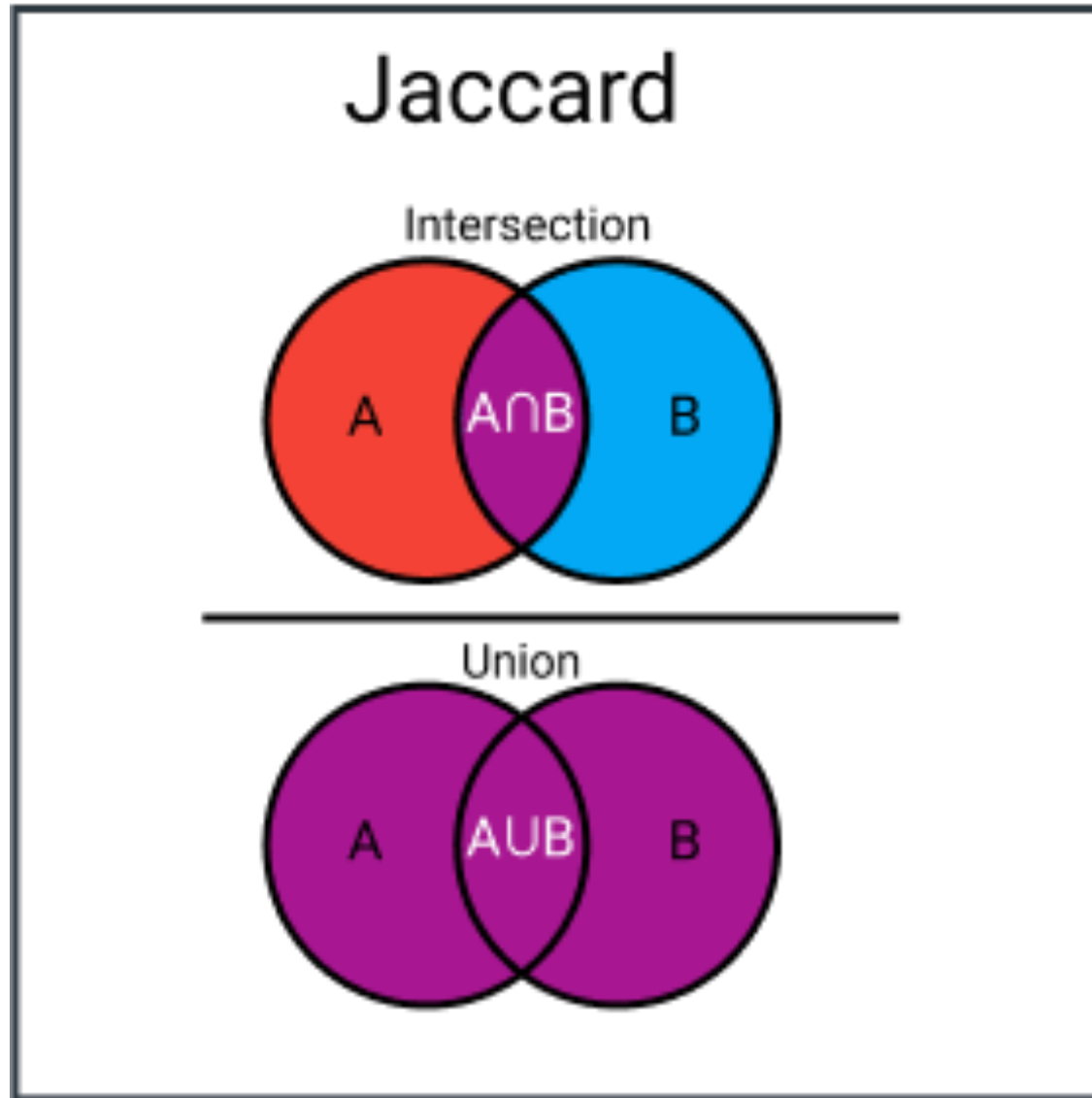
- 'taza', 'casa'
- 'abierto', 'cerrado'
- 'desperdicios', 'deformadores'
- 7589226338, 8572293368
- 8493012576, 8590612473
- 1101000100, 1001110100
- 0100111010, 1101100010

Similitud de Jaccard

- Dados dos conjuntos $\{C^1, C^2\}$, su similitud de Jaccard se define como:

$$J(\mathcal{C}^{(1)}, \mathcal{C}^{(2)}) = \frac{|\mathcal{C}^{(1)} \cap \mathcal{C}^{(2)}|}{|\mathcal{C}^{(1)} \cup \mathcal{C}^{(2)}|} \in [0, 1].$$

Similitud de Jaccard

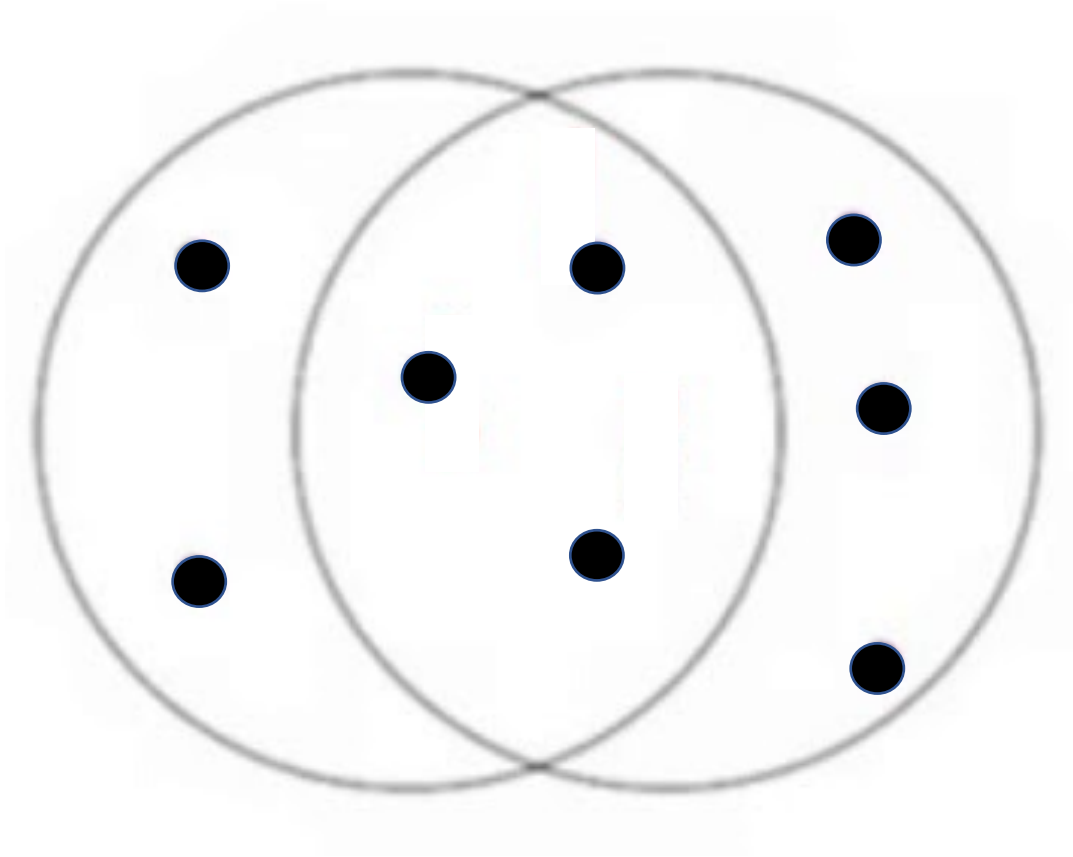


Distancia de Jaccard

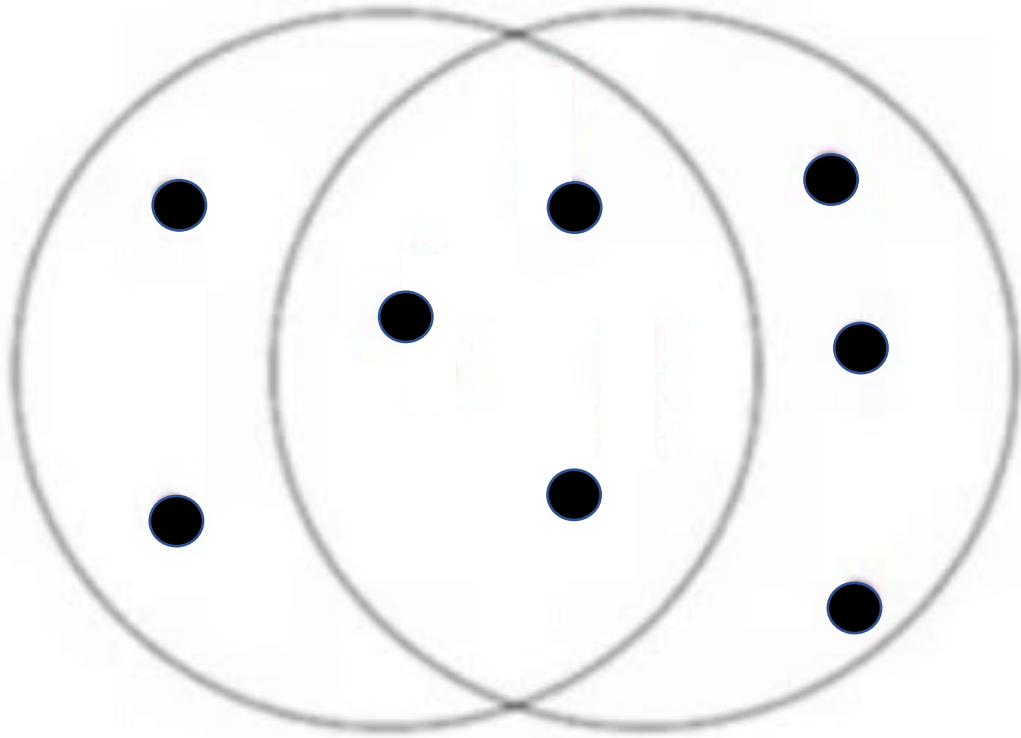
La distancia de Jaccard es:

$$\text{dist}_J(\mathcal{C}^{(1)}, \mathcal{C}^{(2)}) = 1 - J(\mathcal{C}^{(1)}, \mathcal{C}^{(2)})$$

Distancia de Jaccard



Distancia de Jaccard



3 elementos en la
intersección.
8 elementos en la
unión

$$J(A,B) = 3 / 8$$

$$D_j(A,B) = 5 / 8$$

Similitud de Jaccard: Ejercicio

Calcula la similitud de Jaccard de los conjuntos.

- $C^1 : \{1, 6, 3, 7\}$
- $C^2 : \{5, 7, 2, 3\}$

- $C^1 : \{9, 5, 8, 2, 1, 0, 6, 4, 7, 3\}$
- $C^2 : \{4, 6, 2, 3, 12, 15, 13, 2, 5, 16\}$

Traslape de Dos Conjuntos

Número de elementos en común sobre mínimo de elementos de dos conjuntos.

$$ovr(\mathcal{C}^{(1)}, \mathcal{C}^{(2)}) = \frac{|\mathcal{C}^{(1)} \cap \mathcal{C}^{(2)}|}{\min(\mathcal{C}^{(1)}, \mathcal{C}^{(2)})}$$

Traslape de Dos Conjuntos: Ejercicio

Calculo el traslape entre los siguientes pares de conjuntos:

- $C^1 : \{0, 3, 6, 7\}$
- $C^2 : \{2, 3, 5, 7\}$

- $C^1 : \{0, 3, 6, 7\}$
- $C^2 : \{0, 3, 7\}$

- $C^1 : \{2, 3, 7\}$
- $C^2 : \{2, 3, 4, 7\}$

Árboles K – D (*k dimensiones*)

- Árbol binario para realizar búsqueda del vecino más cercano de forma eficiente.
- Cada nivel del árbol se compara con 1 dimensión.

Árboles K – D (*k dimensiones*)

- Para buscar puntos.
 - Se construye el árbol con el conjunto de puntos disponible.
 - Dado un nuevo punto de consulta, se busca el punto más cercano recorriendo el árbol.

Construcción de Árboles K – D (*k dimensiones*)

1. Elige dimensión de forma alternada.

1.1. Por ejemplo para 2D la raíz usará X .

1.2. Sus hijos la dimensión Y

1.3. Los nietos la dimensión X , y así sucesivamente.

2. Inserta punto con valor en la mediana (es posible usar otros criterios para elegir el nodo raíz), de la dimensión seleccionada.

2.1. Puntos menores son descendientes en su rama izquierda.

2.2. Puntos mayores son descendientes en su rama derecha.

Construcción de Árboles K – D (*k dimensiones*)

3. Se repiten los pasos 1 y 2 para todos los descendientes hasta que no haya mas puntos que asignar en el árbol K – D.

Construcción de Árboles K – D (*k dimensiones*): Ejemplo

Supongamos que se tienen los siguientes puntos en un plano 2D:

$$A = (3, 6), (17, 15), (13, 15), (6, 12), \\ (9, 1), (2, 7), (10, 19)$$

Crear un árbol K – D, a partir de A.

Construcción de Árboles K – D (*k dimensiones*): Ejemplo

1. Insertar (3, 6): Dado que el árbol está vacío, conviértalo en el nodo raíz.

2. Insertar (17, 15): Compararlo con el punto del nodo raíz.

Dado que el nodo raíz está alineado con X, el valor de la coordenada X se comparará para determinar si se encuentra en el subárbol derecho o en el subárbol izquierdo.

Este punto estará alineado con Y.

Construcción de Árboles K – D (*k dimensiones*): Ejemplo

3. Insertar (13, 15): El valor X de este punto es mayor que el valor X del punto en el nodo raíz.

Entonces, estará en el subárbol derecho de (3, 6).

Nuevamente compare el valor Y de este punto con el valor Y del punto (17, 15).

Como son iguales, este punto estará en el subárbol derecho de (17, 15).

Este punto estará alineado con X.

Construcción de Árboles K – D (*k dimensiones*): Ejemplo

4. Insertar (6, 12): El valor X de este punto es mayor que el valor X del punto en el nodo raíz.

Entonces, esto estará en el subárbol derecho de (3, 6).

Nuevamente compare el valor Y de este punto con el valor Y del punto (17, 15).

Como $12 < 15$, este punto estará en el subárbol izquierdo de (17, 15).

Este punto estará alineado con X.

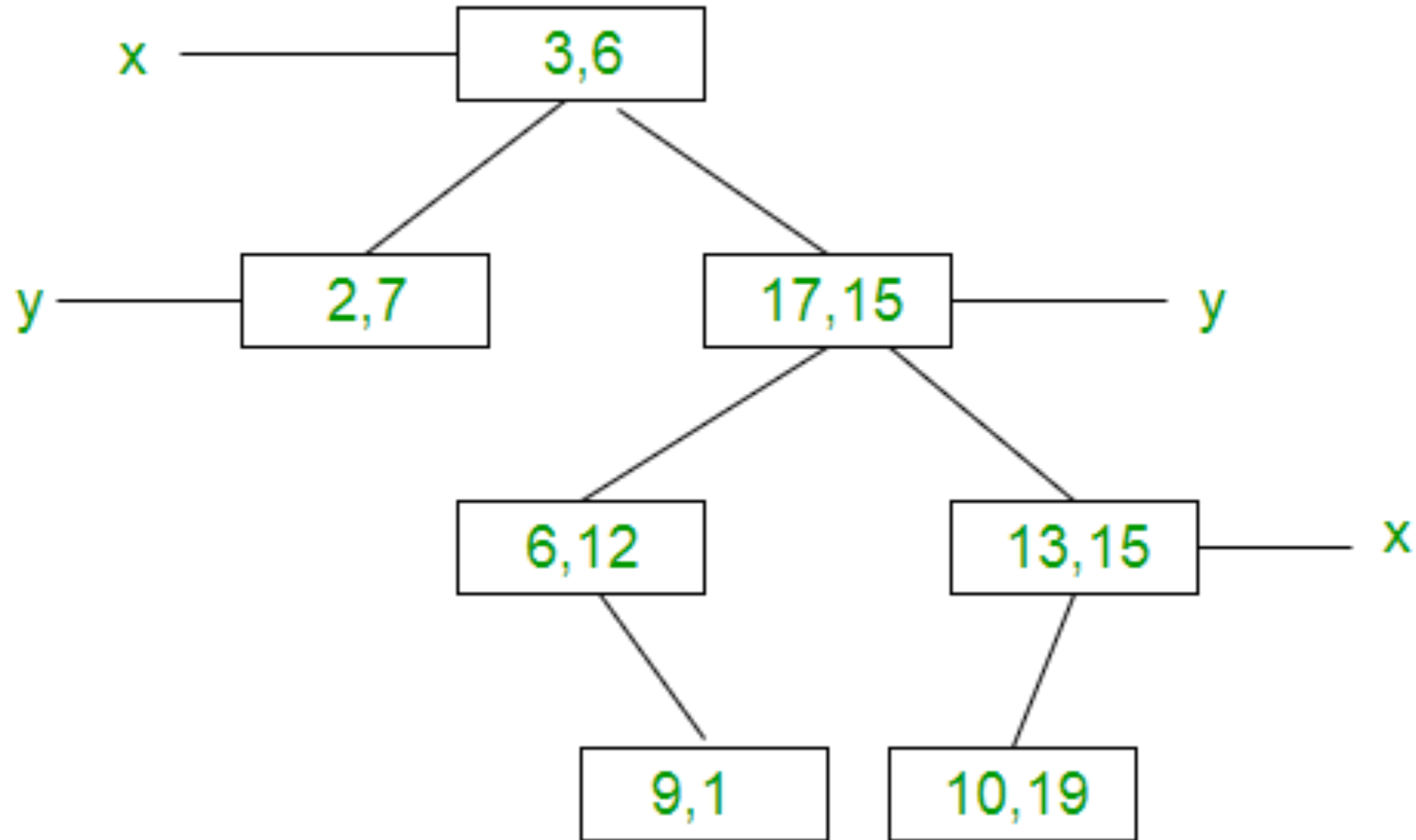
Construcción de Árboles K – D (*k dimensiones*): Ejemplo

5. Insertar (9, 1): De manera similar, este punto estará a la derecha de (6, 12).

6. Insertar (2, 7): De manera similar, este punto estará a la izquierda de (3, 6).

7. Insertar (10, 19): De manera similar, este punto estará a la izquierda de (13, 15).

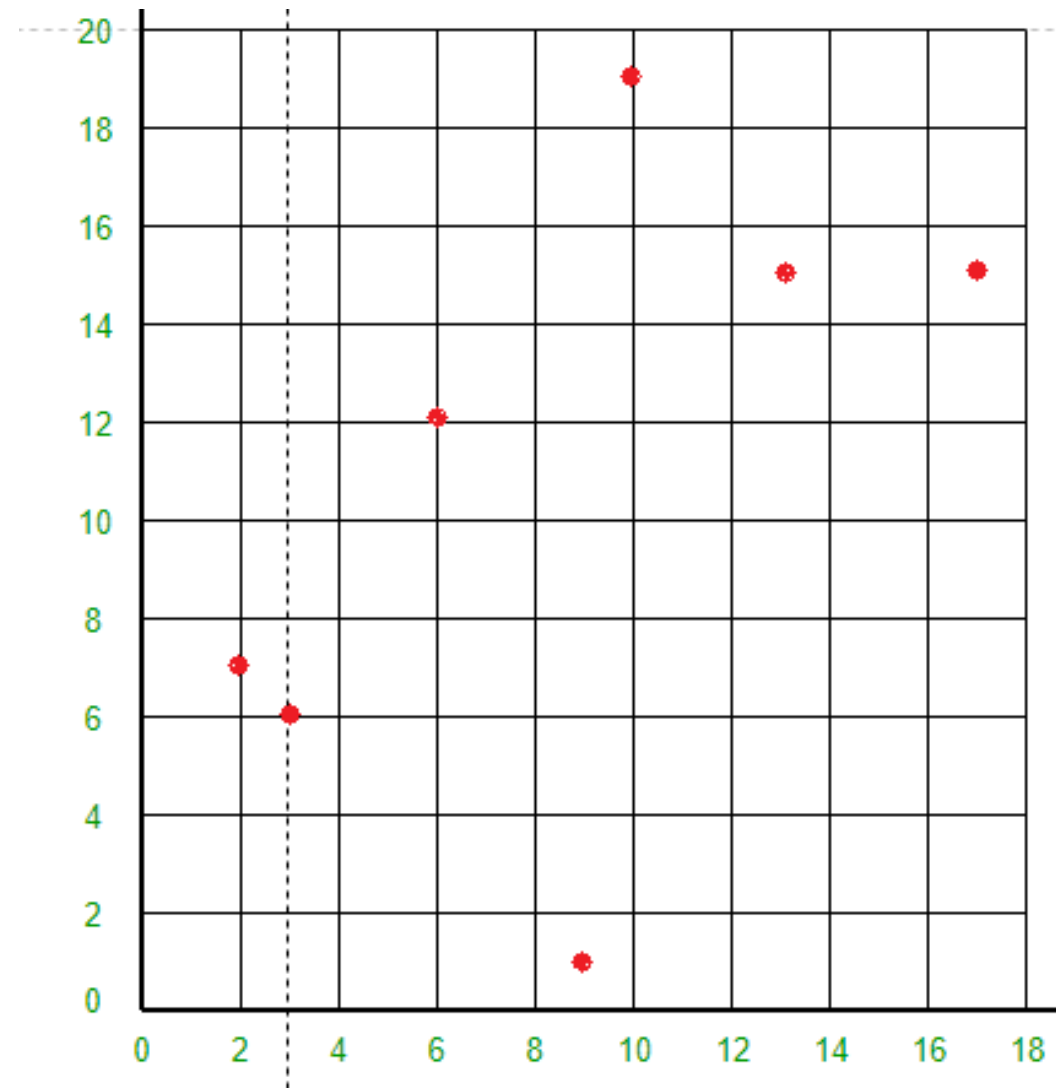
Construcción de Árboles K – D (k dimensiones): Ejemplo



¿Cómo se particiona el espacio 2D con los 7 puntos del árbol?

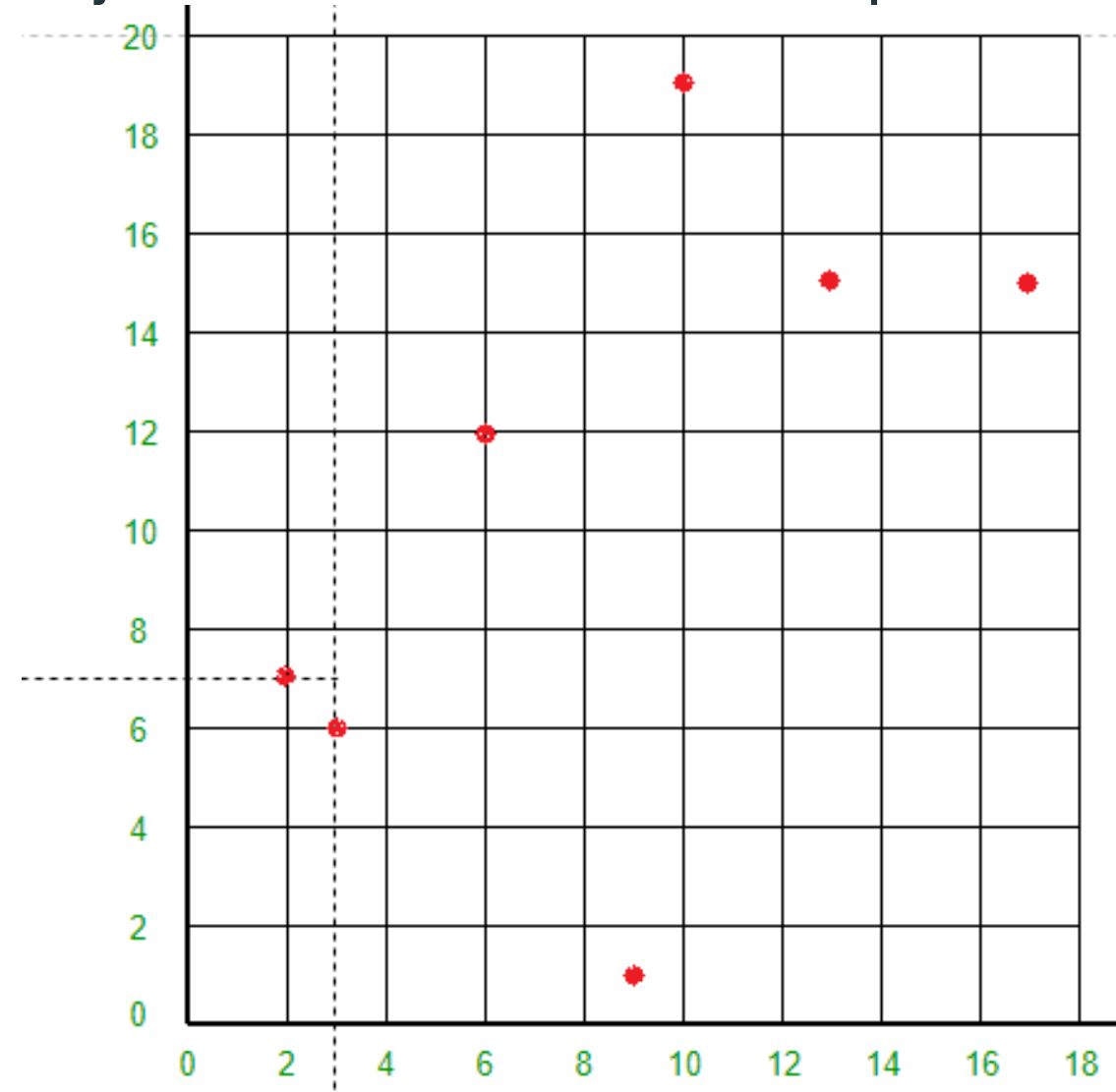
Partición del Espacio (Árboles K – D)

Punto (3, 6). Dibujar línea en $X = 3$.



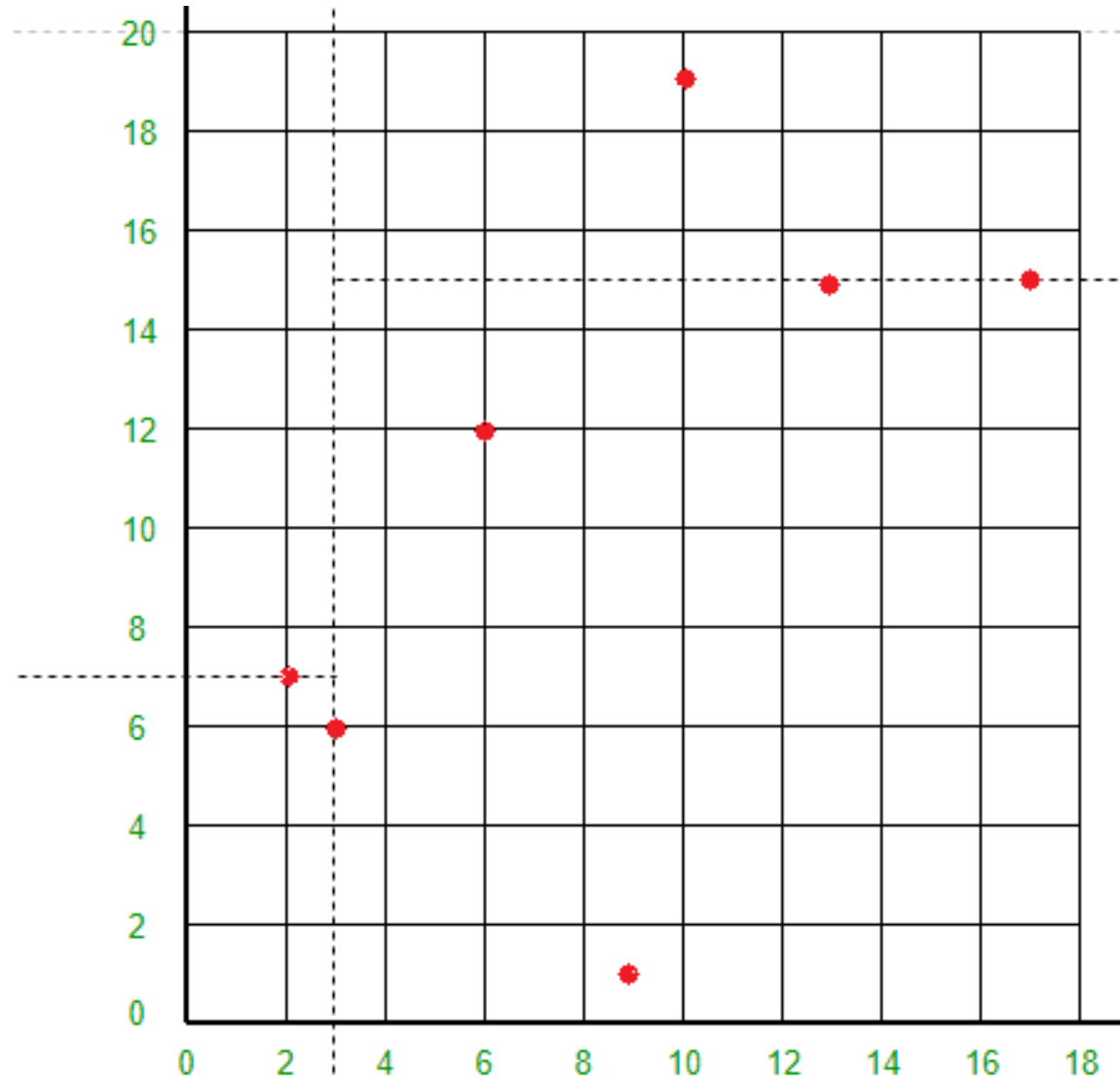
Partición del Espacio (Árboles K – D)

Punto (2, 7). Dibujar línea en $Y = 7$ a la izquierda de la línea $X = 3$.



Partición del Espacio (Árboles K – D)

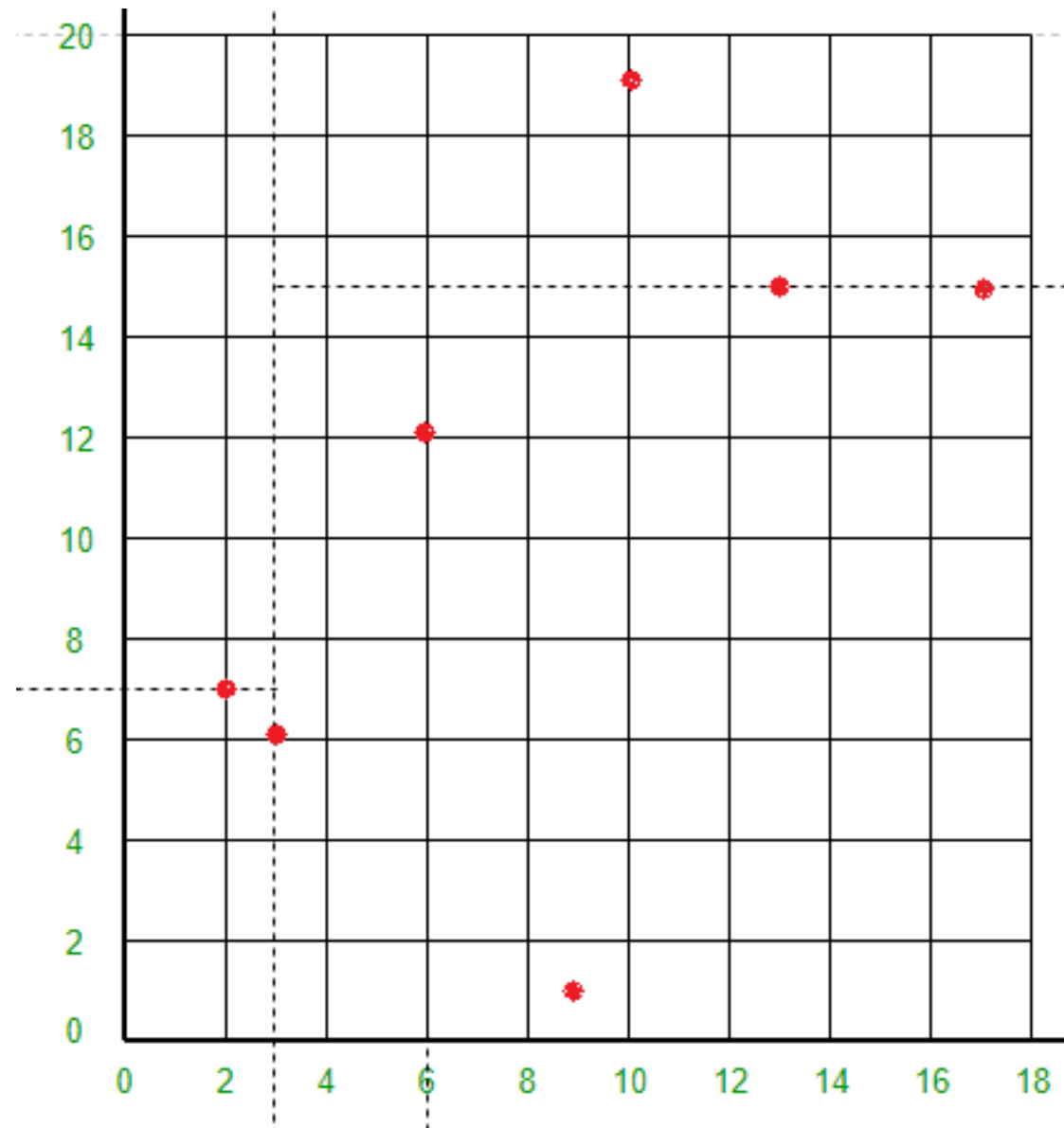
Punto (17, 15). Dibujar línea en $Y = 15$ a la derecha de la línea $X = 3$



Partición del Espacio (Árboles K – D)

Punto (6, 12).

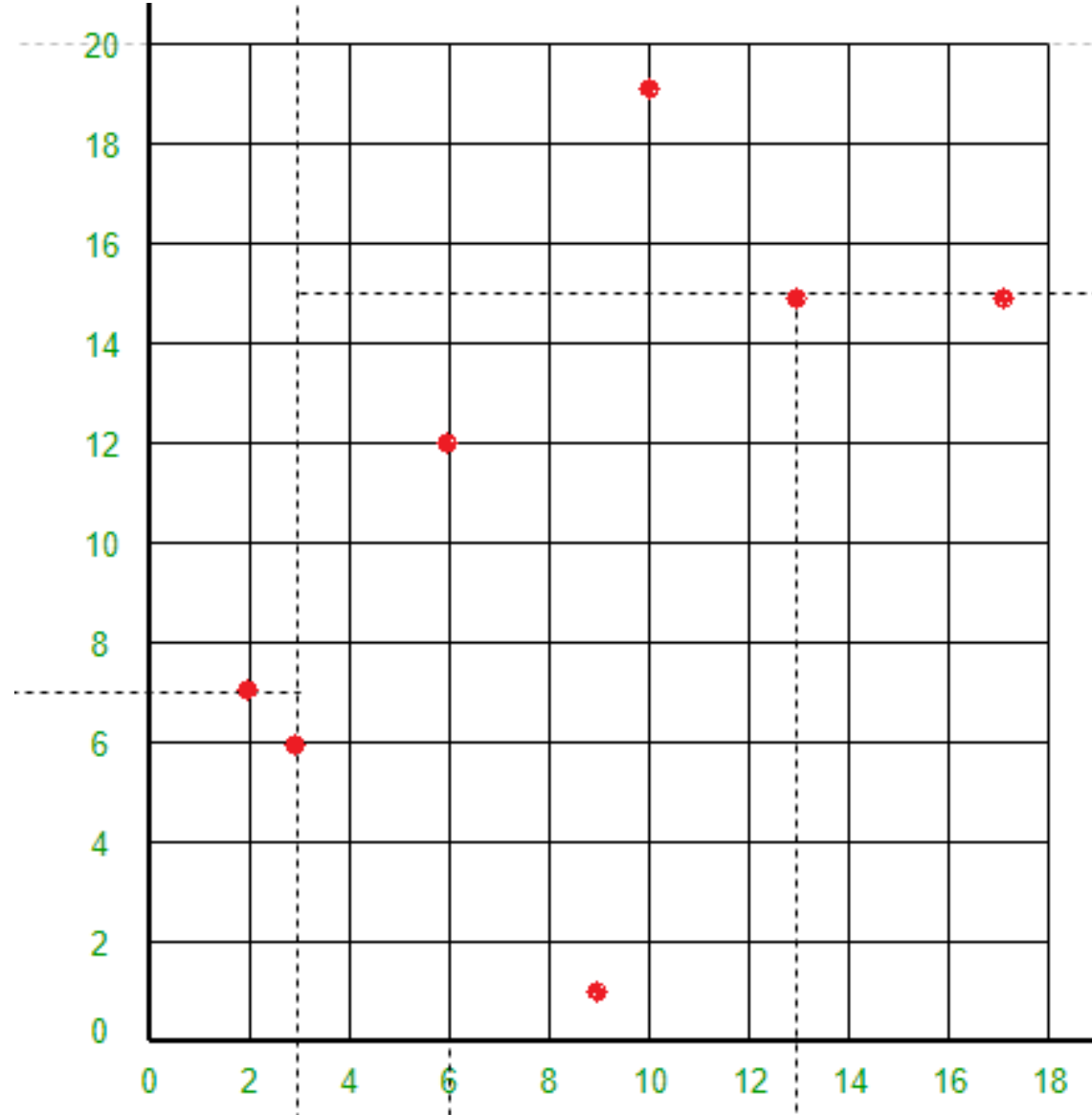
Dividirá el espacio debajo de la línea $Y = 15$ y a la derecha de la línea $X = 3$ en dos partes. Dibujar línea en $X = 6$ a la derecha de la línea $X = 3$ y debajo de la línea $Y = 15$



Partición del Espacio (Árboles K – D)

Punto (13, 15).

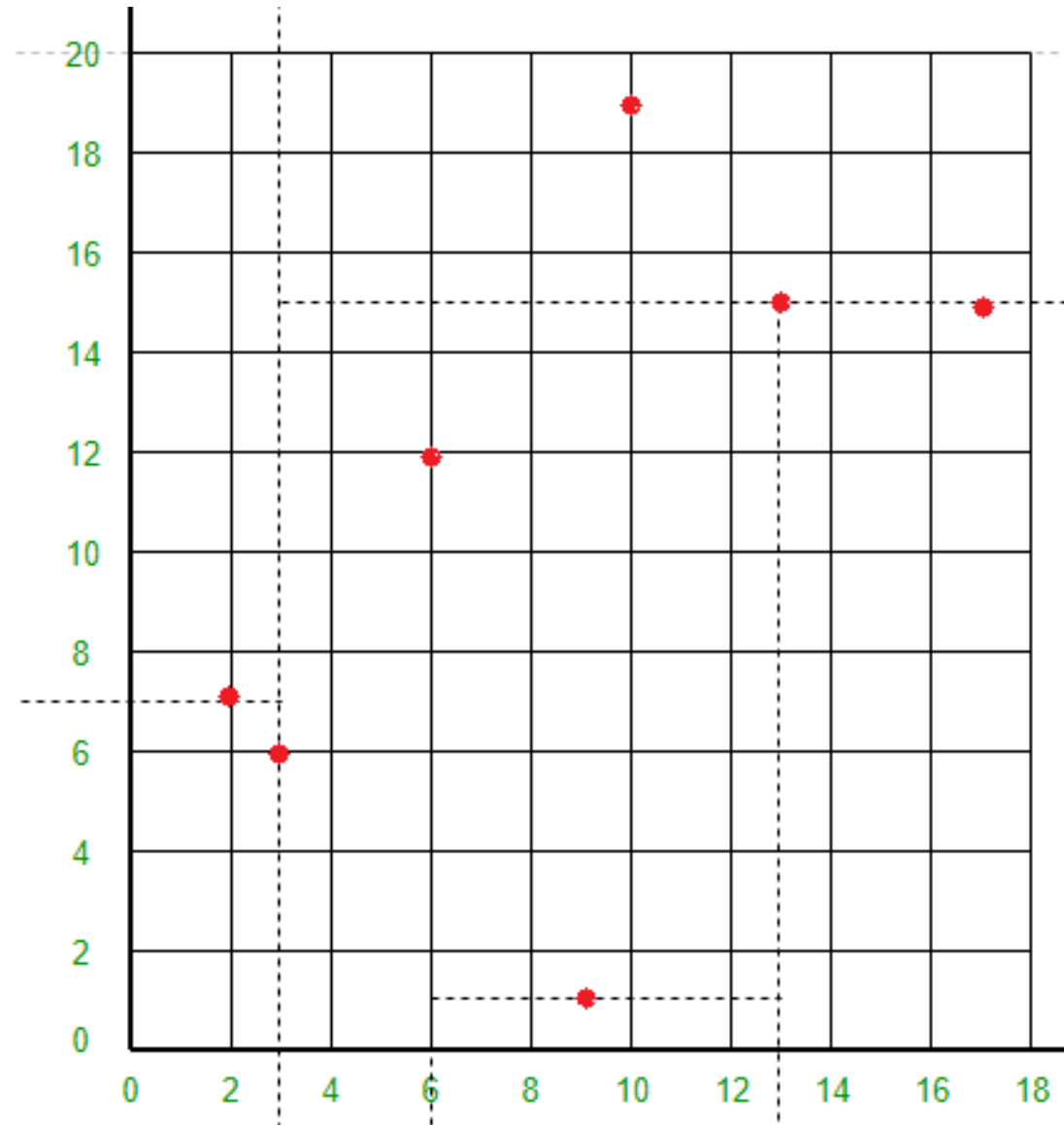
Dividirá el espacio debajo de la línea $Y = 15$ y a la derecha de la línea $X = 6$ en dos partes. Dibujar línea en $X = 13$ a la derecha de la línea $X = 6$ y debajo de la línea $Y = 15$



Partición del Espacio (Árboles K – D)

Punto (9, 1).

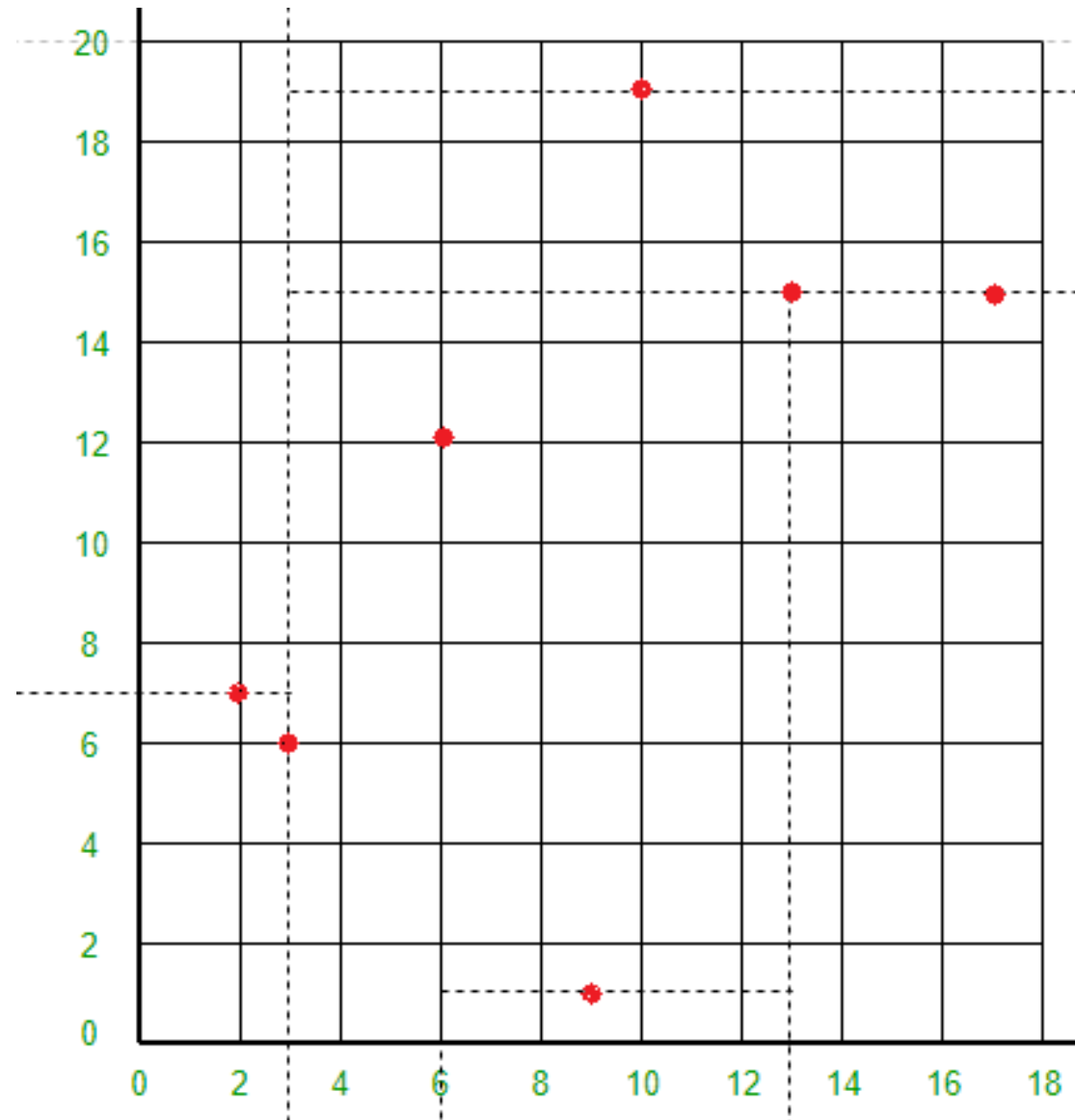
Dividirá el espacio entre las líneas $X = 3$, $X = 6$ e $Y = 15$ en dos partes. Dibujar línea en $Y = 1$ entre las líneas $X = 3$ y $X = 13$.



Partición del Espacio (Árboles K – D)

Punto (10, 19).

Dividirá el espacio a la derecha de la línea $X = 3$ y arriba de la línea $Y = 15$ en dos partes. Dibujar línea en $Y = 19$ a la derecha de la línea $X = 3$ y arriba de la línea $Y = 15$



Crear el árbol binario con los siguientes puntos:

$$A = \{(2, 3), (5, 4), (9, 6), (4, 7), (8, 1), (7, 2)\}$$

Tomando como nodo raíz la mediana = (7, 2)

Partición del Espacio (Árboles K – D): Ejercicio

Crear el árbol binario con los siguientes puntos:

$$A = \{(2, 3), (5, 4), (9, 6), (4, 7), (8, 1), (7, 2)\}$$

Generar la partición del espacio.

Árboles K – D: Inserción de Puntos al Árbol

1. Recorre el árbol a partir de la raíz y moviéndose hacia el descendiente correspondiente.
2. Cuando se encuentra el nodo padre del punto a insertar, se agrega a la derecha o izquierda dependiendo del valor en la dimensión de partición.

3. En caso de estar desbalanceado, se aplica un algoritmo de re-balanceo para evitar pérdida de rendimiento

Árboles K – D: Búsqueda del Vecino Más Cercano

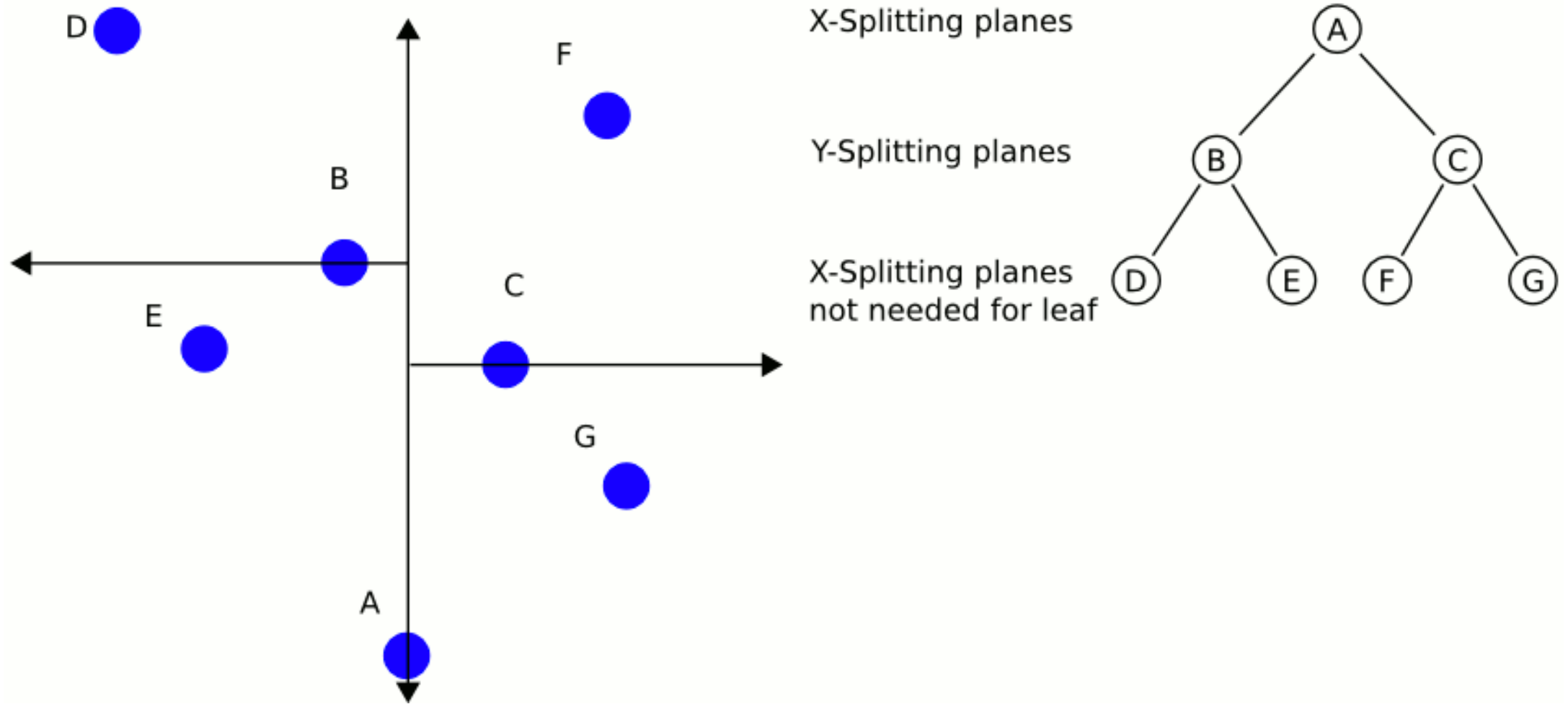
Recorre el árbol a partir de la raíz y moviéndose hacia el descendiente correspondiente.

1. Mantén el punto más cercano c_{\min} y quita los nodos del árbol que están más alejados a este.
2. Recorre los sub-árboles restantes.
 1. Existen heurísticas para elegir aquel que permita quitar más nodos.

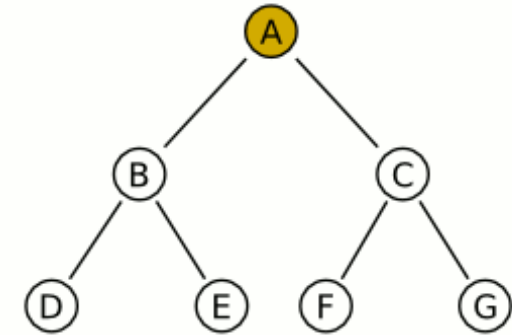
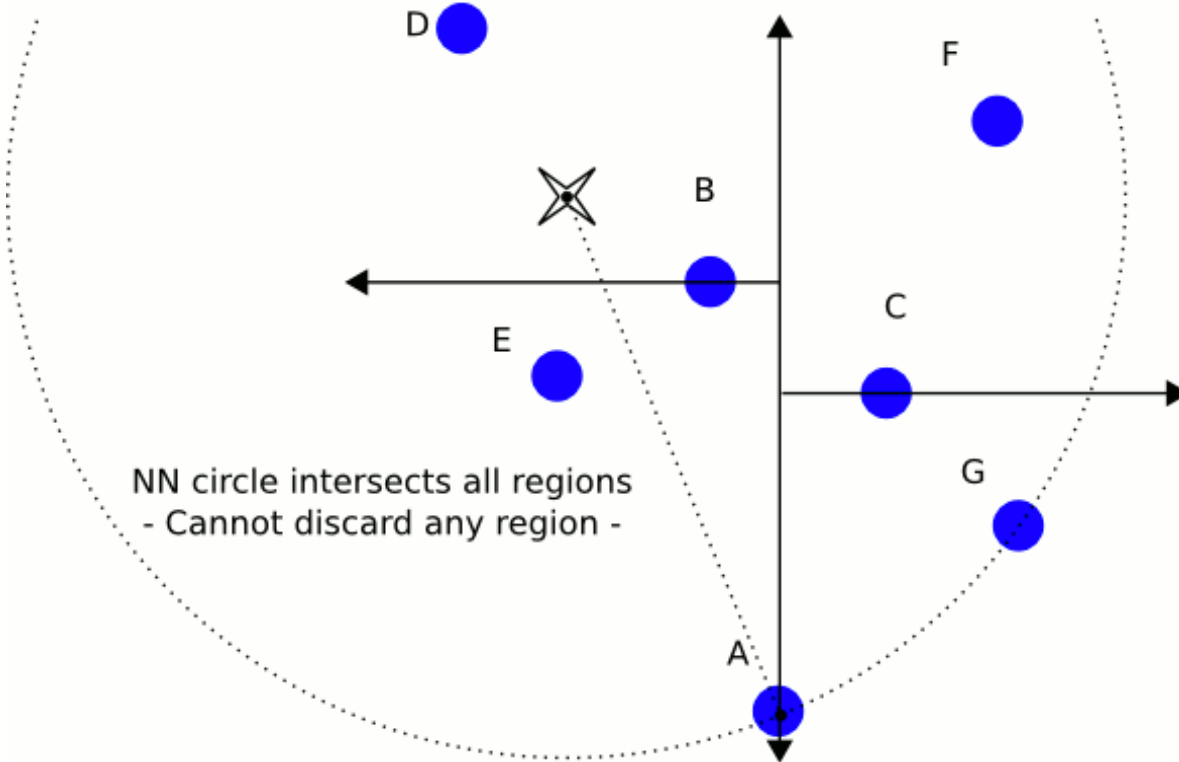
Árboles K – D: Búsqueda del Vecino Más Cercano (Complejidad)

- En el peor de los casos el tiempo de búsqueda es $O(n)$.
- Pero en promedio es $O(\log(n))$
- Algoritmo sufre por la maldición de la dimensionalidad.

Árboles K – D: Ejemplo de Búsqueda

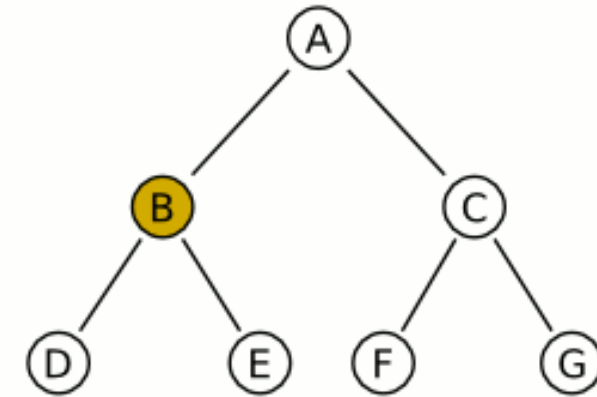
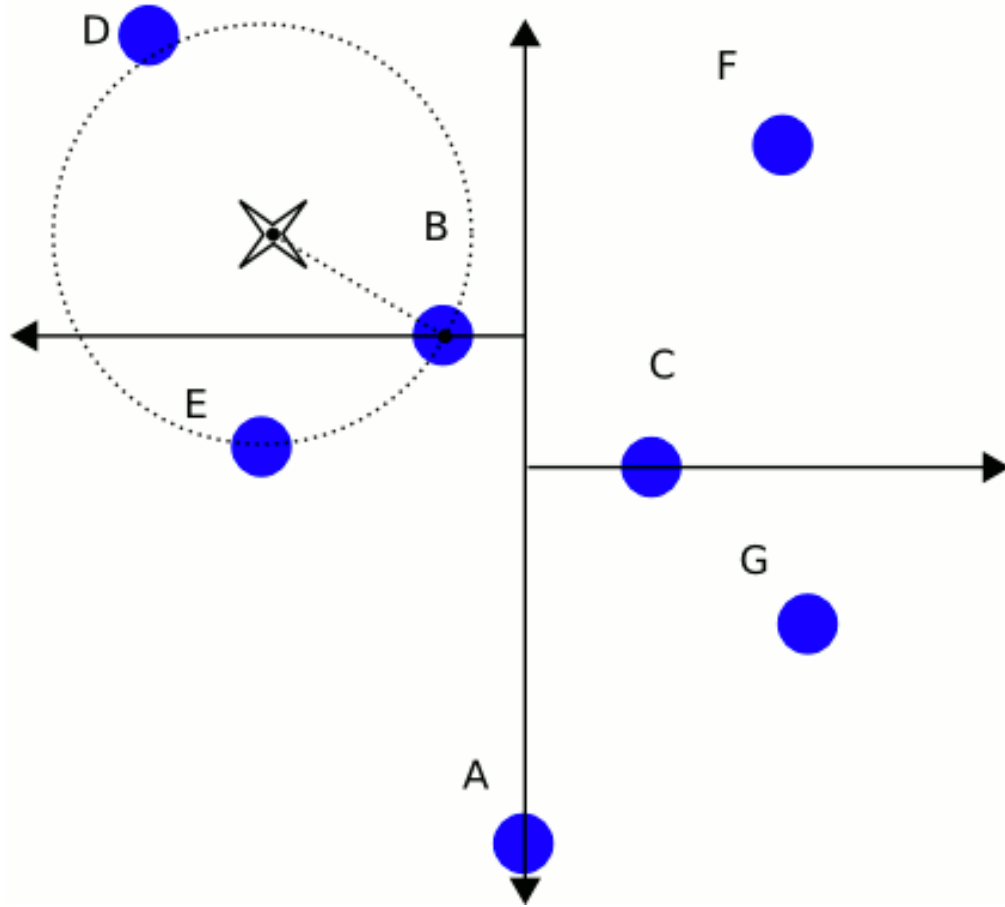


Árboles K – D: Ejemplo de Búsqueda



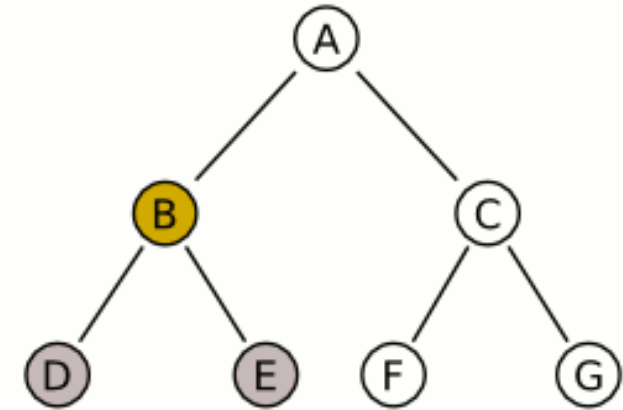
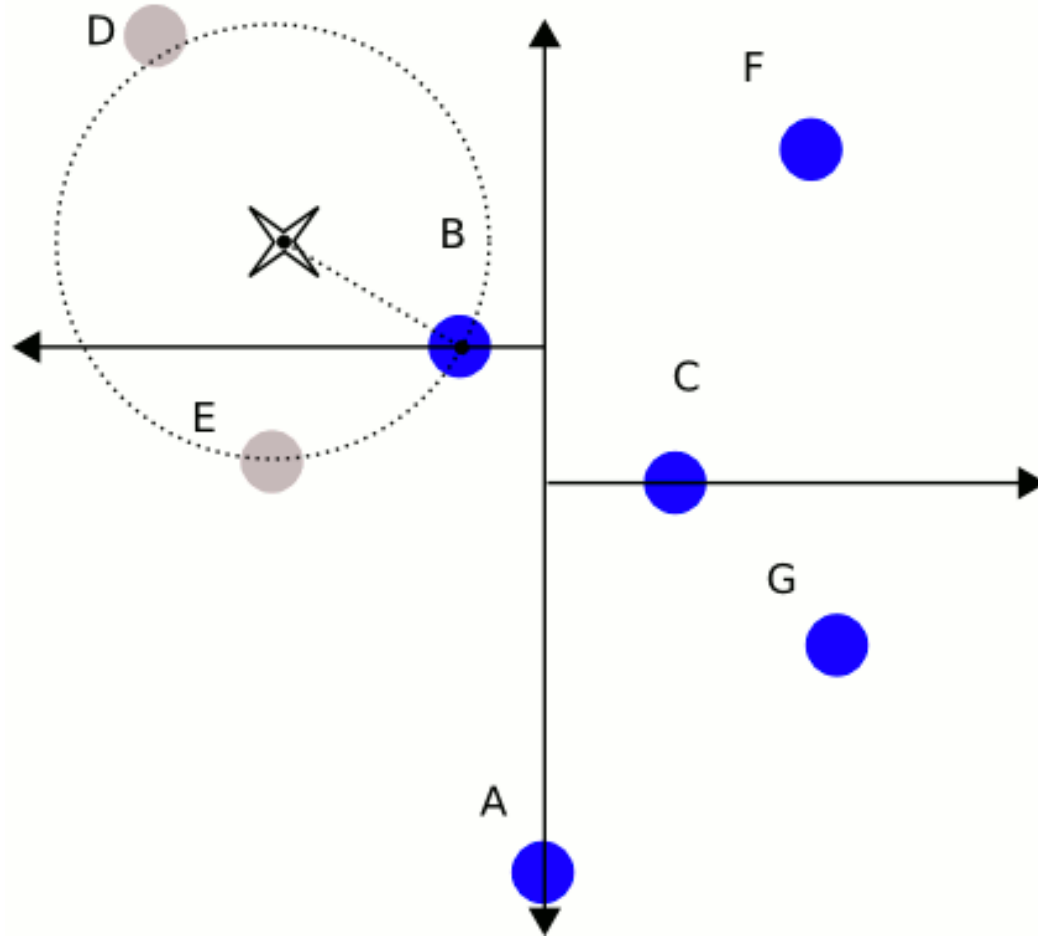
Start at A, then proceed in depth-first search (maintain a stack of parent-nodes if using a singly-linked tree). Set best estimate to A's distance. Then examine left child node

Árboles K – D: Ejemplo de Búsqueda



Calculate B's distance and compare against best estimate
- It is smaller distance, so update best estimate. Examine children (left then right)

Árboles K – D: Ejemplo de Búsqueda



D & E Discarded as B
(already visited) is closer.
B is the best estimate for B's sub-branch
Proceed back to parent node

Árboles K – D, Ejemplo de Búsqueda: Ejercicio

Agregar el siguiente punto a A:

$\{(8, 7)\}$

Árboles K – D, Ejemplo de Búsqueda: Ejercicio

Después, busca los vecinos más cercanos en A de los siguientes puntos:

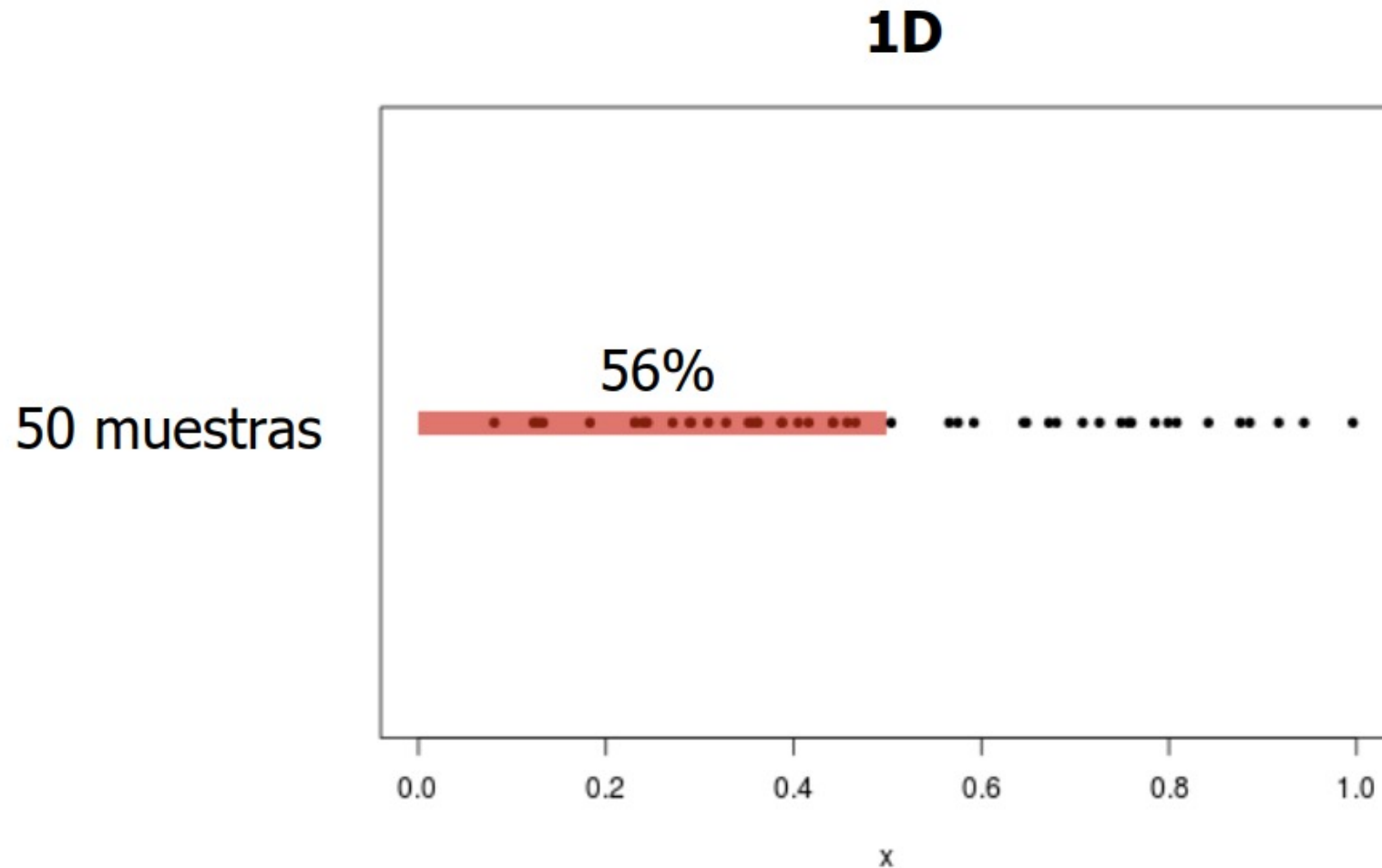
$\{(7, 2),$

$(5, 4),$

$(9, 6)\}$

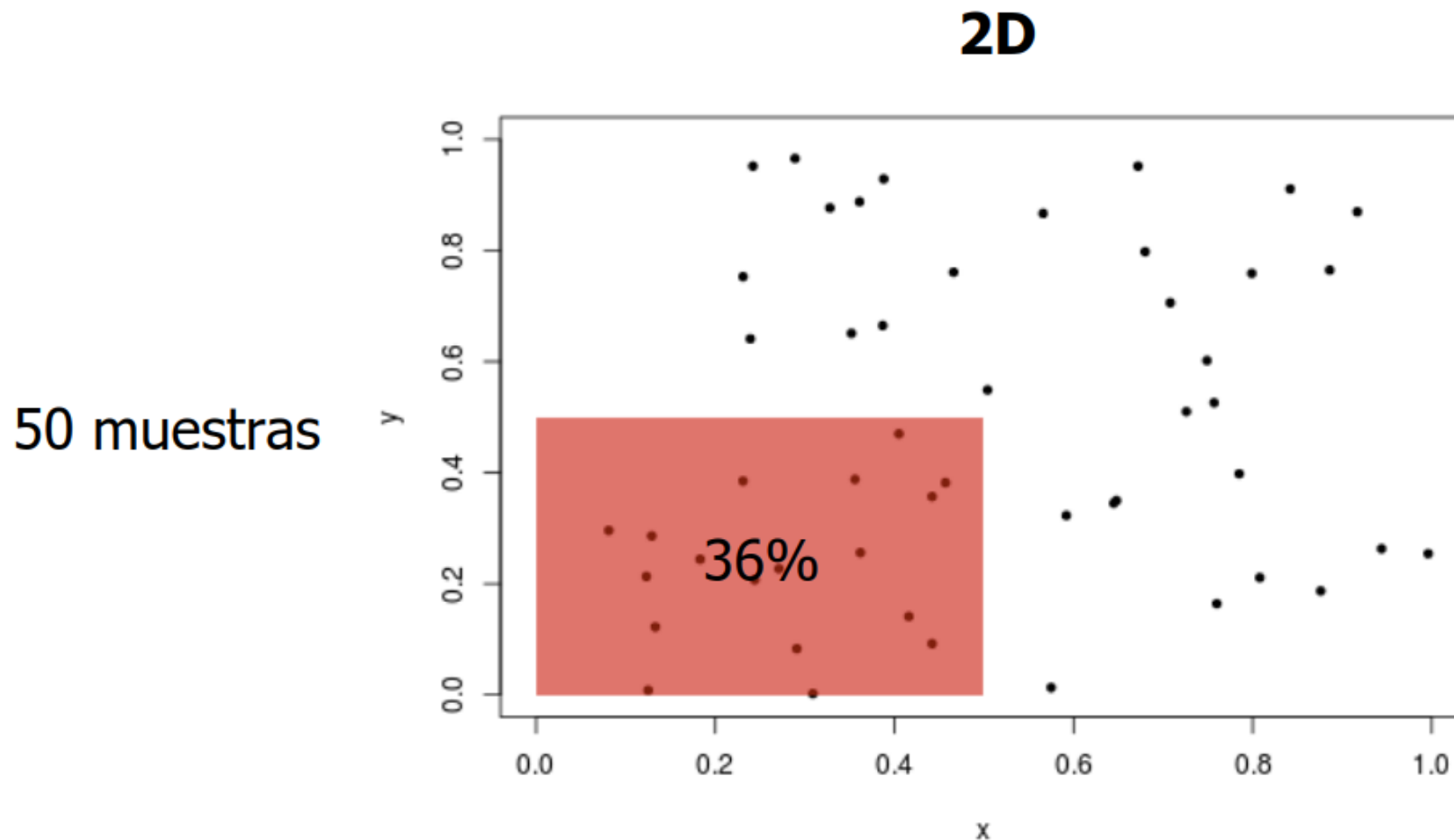
Dimensionalidad (La Maldición)

Objetos cada vez más dispersos conforme aumenta el número de dimensiones.



Dimensionalidad (La Maldición)

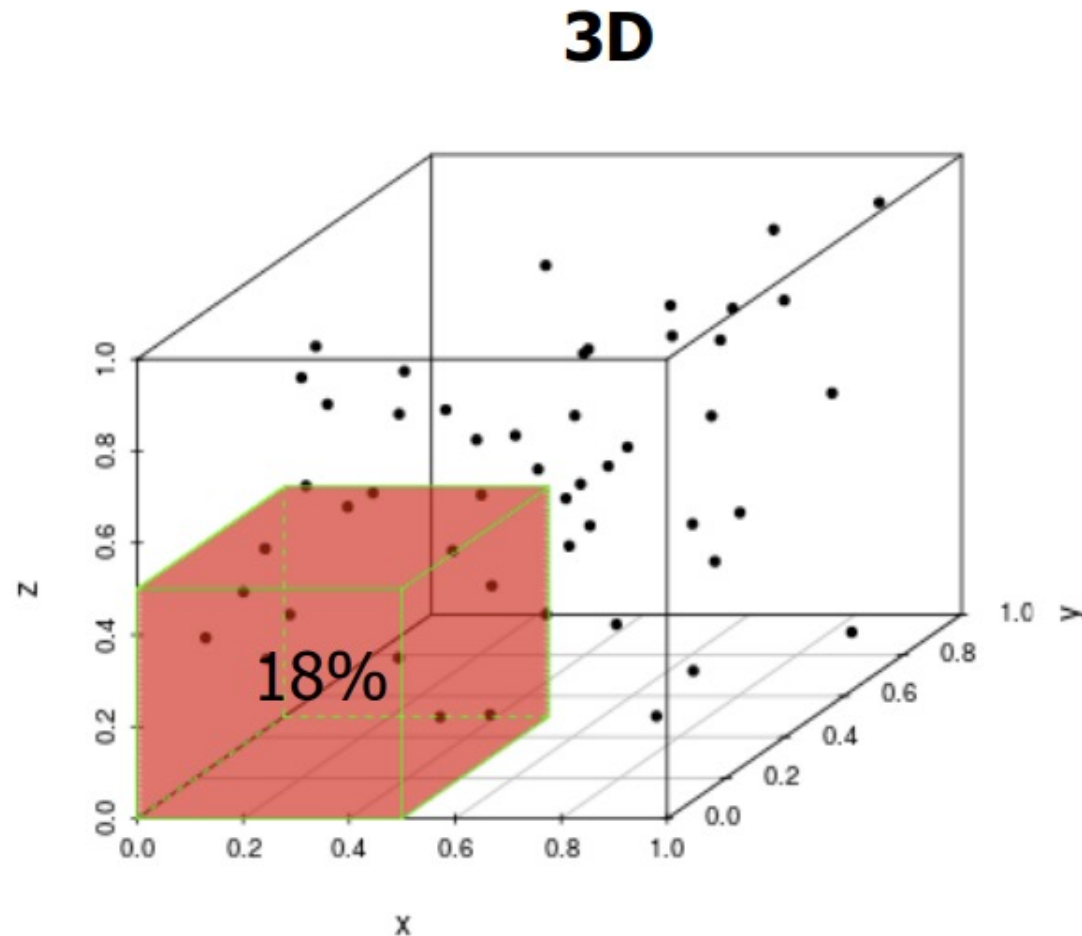
Objetos cada vez más dispersos conforme aumenta el número de dimensiones.



Dimensionalidad (La Maldición)

Objetos cada vez más dispersos conforme aumenta el número de dimensiones.

50 muestras



Matriz Documento – Término

Documentos

D_1 = Él duerme con su hijo mientras su perro duerme

D_2 = Ella duerme de día y su perro duerme de noche

	duerme	perro	hijo	mientras	noche	día	el	ella	de	con	su	y
D_1	1	1	1	1	0	0	1	0	0	1	1	0
D_2	1	1	0	0	1	1	0	1	1	0	1	1

Matriz Documento – Término

	duerme	perro	hijo	mientras	noche	día	el	ella	de	con	su	y
D ₁	1	1	1	1	0	0	1	0	0	1	1	0
D ₂	1	1	0	0	1	1	0	1	1	0	1	1

Bolsa de palabras

D₁ = {duerme, perro, hijo, mientras, él, con, su}

D₂ = {duerme, perro, noche, día, ella, de, su, y}

Conjunto de documentos.

D_1 = Él duerme con su hijo mientras su perro duerme

D_2 = Ella duerme de día y su perro duerme de noche

Matriz Documento – Término

D_1 = Él duerme con su hijo mientras su perro duerme

w_1

w_3

w_2

w_1

D_2 = Ella duerme de día y su perro duerme de noche

w_1

w_2

w_1

	duerme	perro	hijo
D_1	1	1	1
D_2	1	1	0



$D_1 = \{1, 2, 3\}$

$D_2 = \{1, 2\}$



Bolsa de
palabras
binaria.

Matriz Documento – Término

D_1 = Él duerme con su hijo mientras su perro duerme

w_1

w_3

w_2

w_1

D_2 = Ella duerme de día y su perro duerme de noche

w_1

w_2

w_1

	duerme	perro	hijo
D_1	1	1	1
D_2	1	1	0



$D_1 = \{1, 2, 3\}$

$D_2 = \{1, 2\}$



Bolsa de palabras binaria.

	w_1	w_2	w_3
D_1	2	1	1
D_2	2	1	0



$D_1 = \{1, 1, 2, 3\}$

$D_2 = \{1, 1, 2\}$



Bolsa de palabras con frecuencia.

Matriz Documento – Término: Aplicaciones

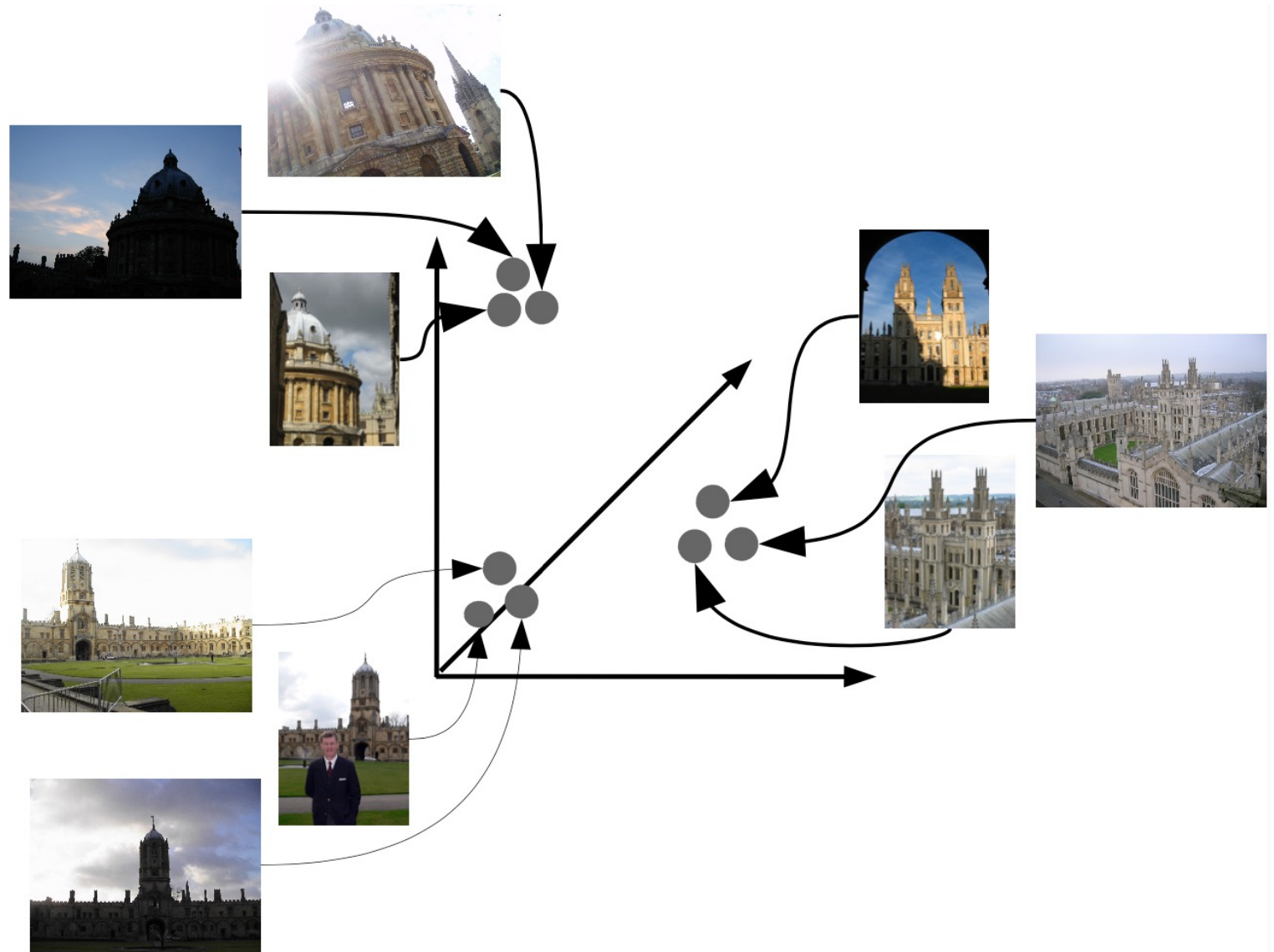
Finding the topic
and the main idea



*How Google Search
continues to improve results*

Representando Imágenes

Buscamos mapear las imágenes a una representación compacta, discriminativa, descriptiva, robusta y rápida de obtener.



Dos tareas fundamentales:

1. Detección de regiones de interés.
 2. Descripción de cada región.
- Imagen – conjunto de vectores característicos.

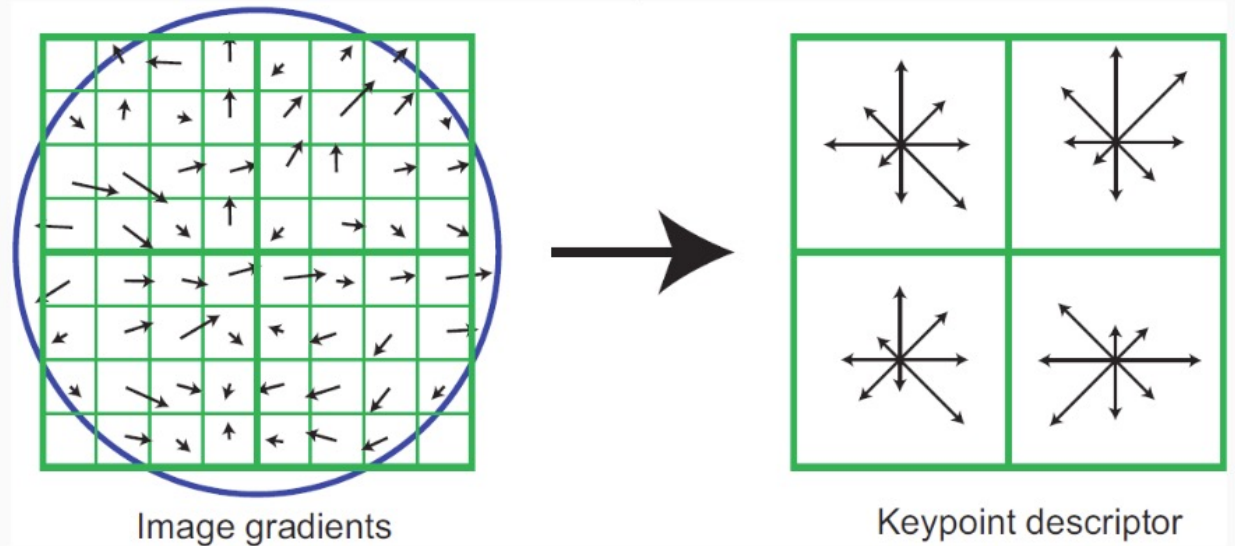
Representando Imágenes

- Imagen – conjunto de vectores característicos.

Detección

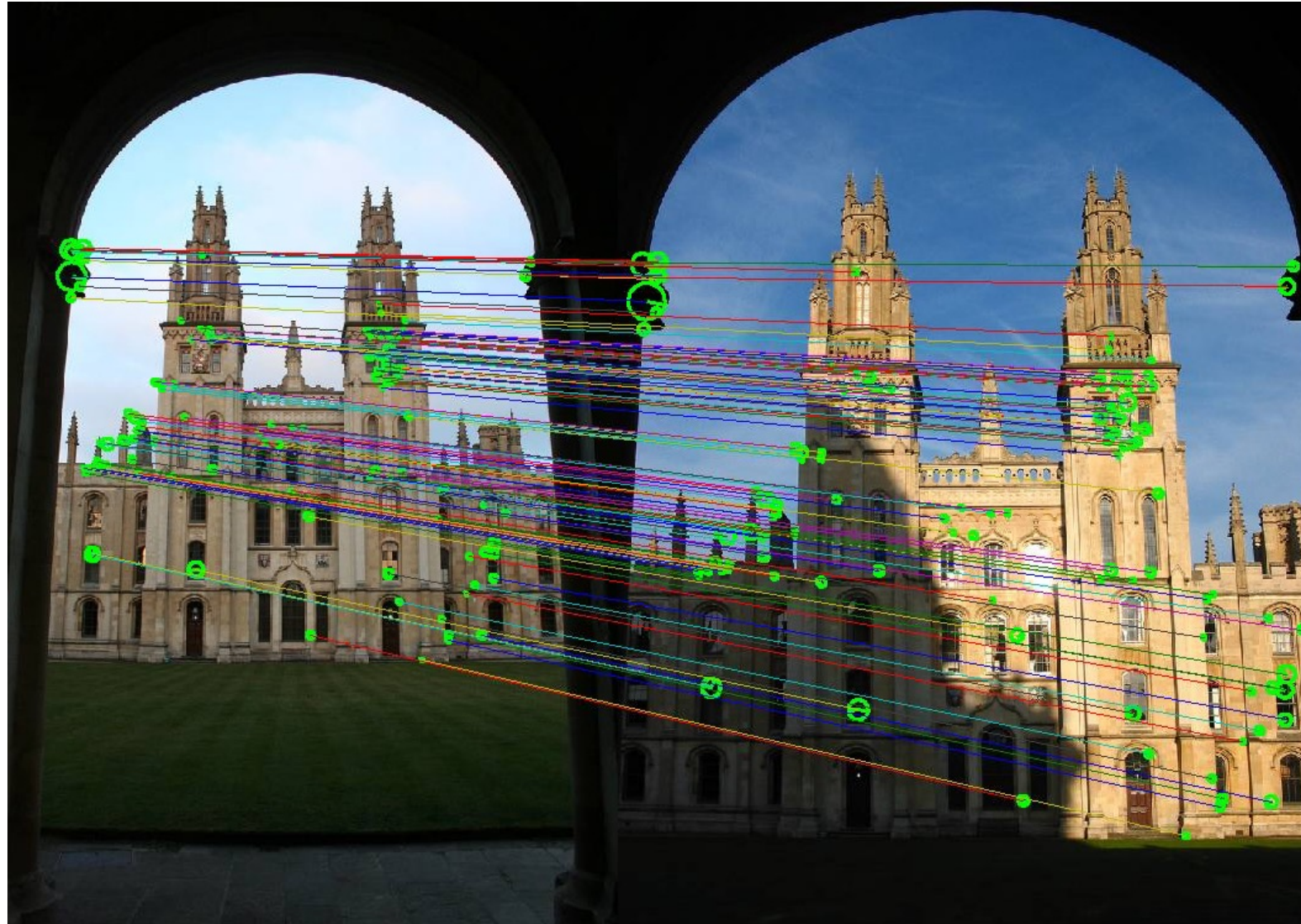


Descripción

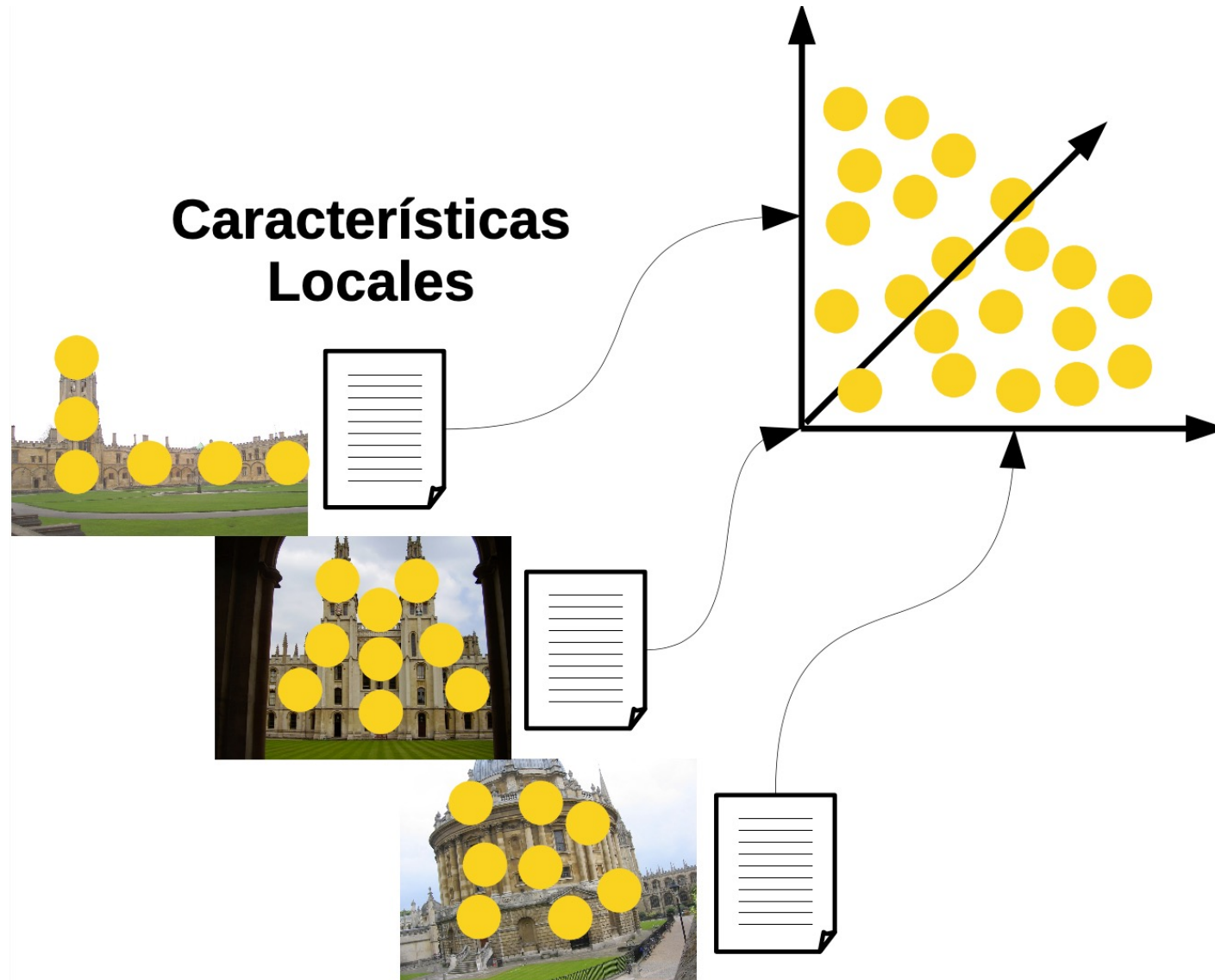


¿Cómo Comparar Imágenes?

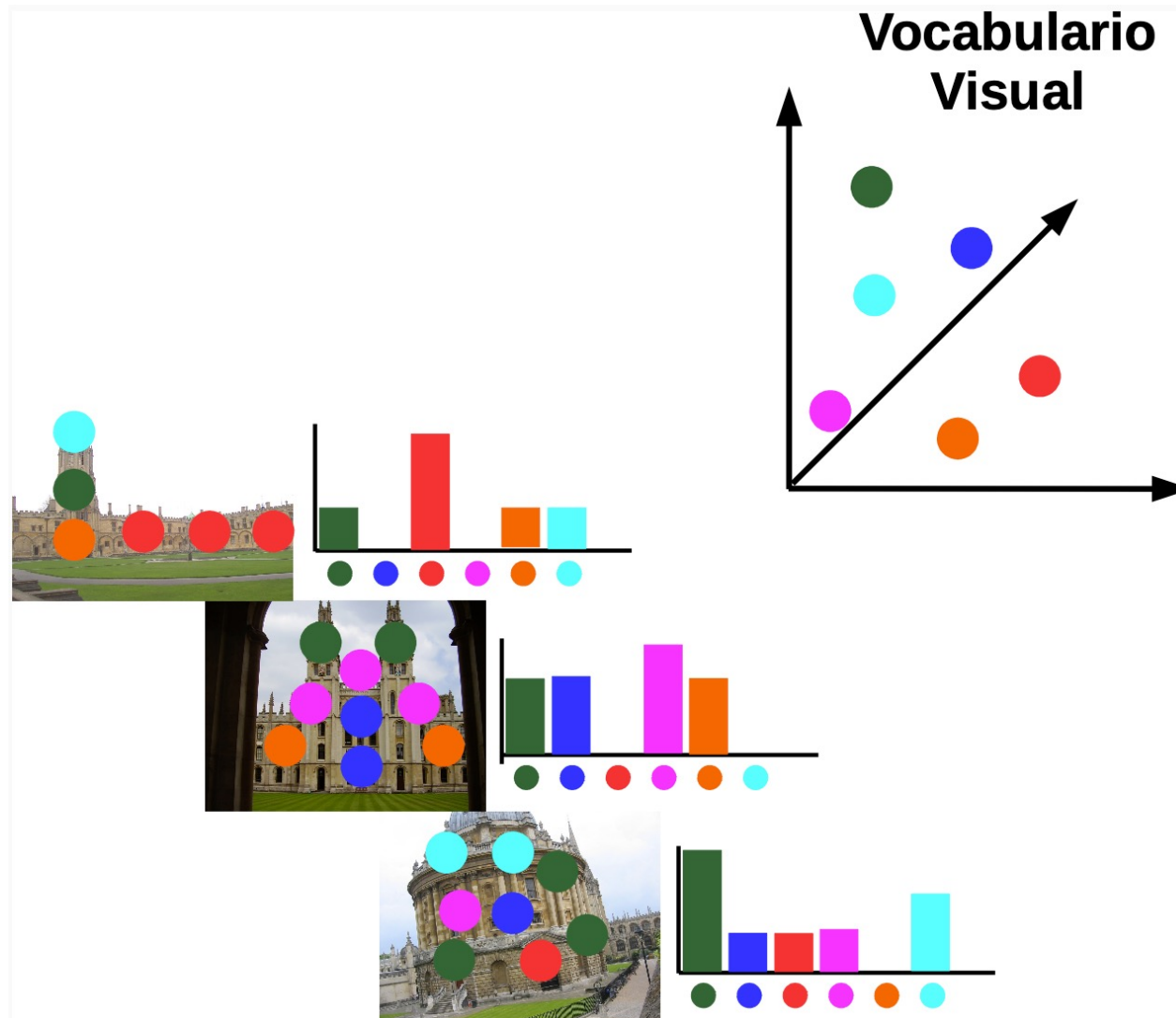
Búsqueda de características similares.



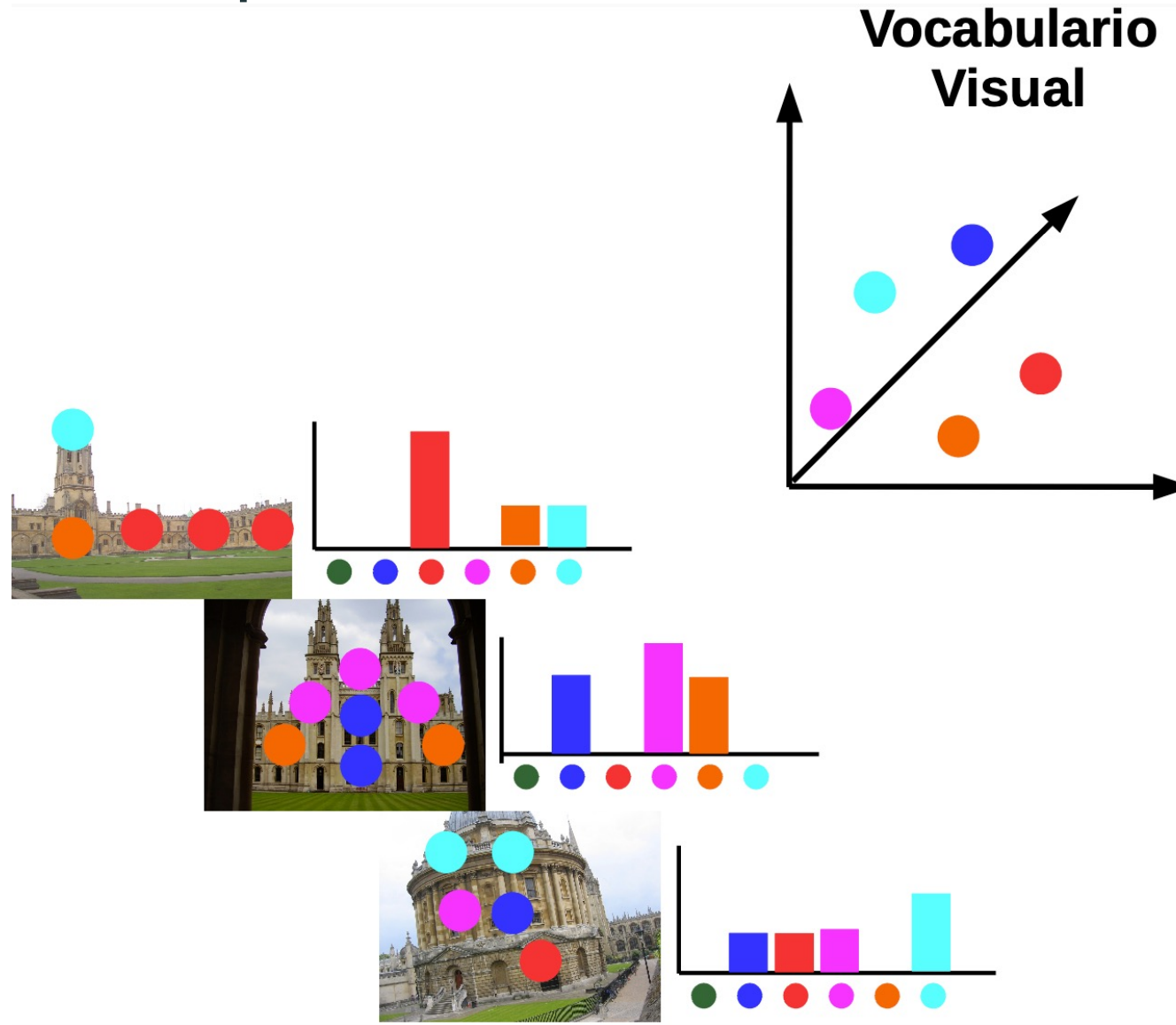
Palabras – Características Locales.



Stemming – Cuantización (por ejemplo: K-Means)



Stop Words – Removerlas



¿Cómo Buscar Imágenes / Documentos Similares?

- Filas y columnas de la matriz documento – término usualmente están dispersas y se representan por conjuntos o bolsas.
- Compara solo las/los que compartan al menos una característica/palabra.
- Se ordenan por valor de distancia o similitud.

Palabra	Ocurrencia
0	1, 4
1	9, 11, 13
2	3, 7, 8, 12, 15
3	5, 6
4	2, 4, 7, 12
5	6, 8, 11
...	...

○ Búsqueda por índice inverso

- Recupera los conjuntos o bolsas de documentos donde ocurren las palabras en D .
- Calcula la distancia o similitud entre D y cada elemento en la lista.
- Ordena d de acuerdo a las distancias o similitudes calculadas.

Son una secuencia de n objetos, que pueden ser:

- Símbolos (n – gramas de símbolos)
- Palabras (n – gramas de palabras)

Ejercicio:

Genera los 2 – gramas y 3 – gramas de símbolos y palabras de la siguiente oración:

Ella toma café y él toma mate.

- Se utilizan en uno de los más exitosos modelos de lenguaje para el reconocimiento de voz.
- En los editores de textos para recomendar cual va a ser la palabra siguiente o para detectar posibles errores de ortografía.
- Son utilizados comúnmente como base para el análisis estadístico de texto.