

# Técnicas de Clustering



# Programa

- Introducción
- Métodos Divisivos
- Métodos Jerárquicos
- Algunos otros métodos
- Cuantos clusters? Gap y Estabilidad



# Introducción





# Introducción



# Introducción

- Definiciones previas:
  - Cluster: Agrupamiento de objetos.
  - Idea de grupo: Objetos que son similares entre sí pero diferentes del resto.
  - Métrica: medida de similitud entre objetos

# Idea intuitiva

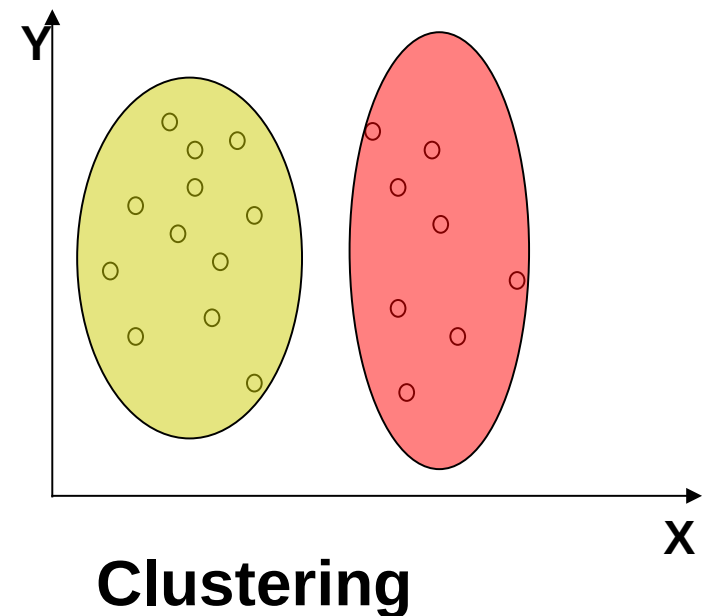
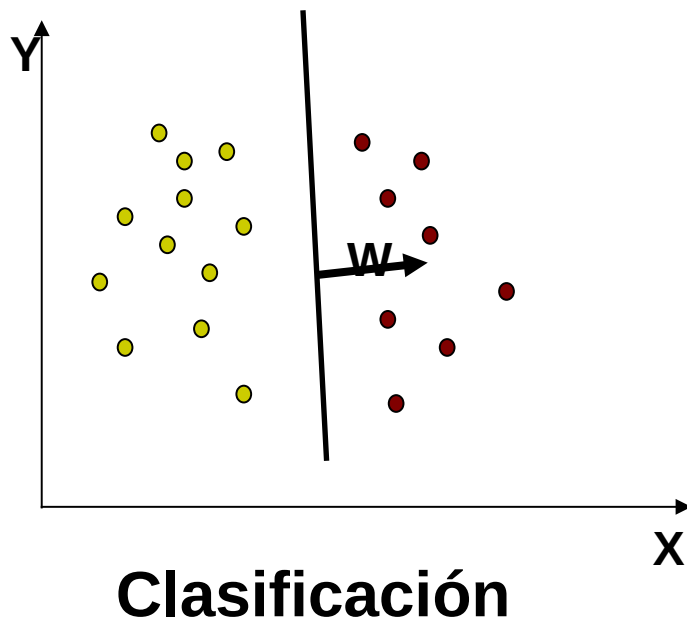
- Datos
  - un conjunto de objetos (datos)
  - una medida de similitud entre ellos (métrica)
- Encontrar una partición de los mismos /
  - Mismo grupo → Similares
  - Distinto grupo → Distintos
  - Que tenga sentido, que sea interesante

# Objetivos

- Descubrir información
  - Encontrar “grupos naturales” en un conjunto de datos del que no se conocen “clases”.
  - Encontrar jerarquías de similaridad en los datos (taxonomías)
- Resumir los datos
  - Encontrar “prototipos” que sean representativos de un conjunto grande de ejemplos
- Dividir los datos + otros...

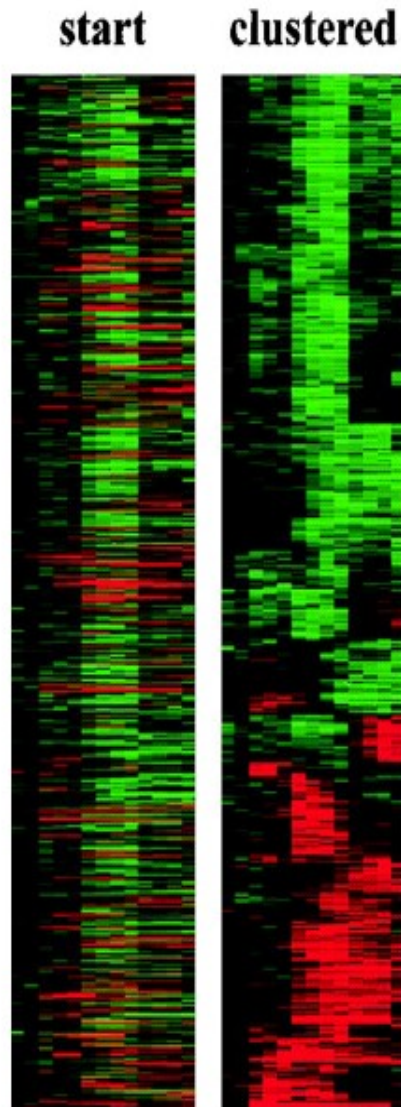
# Clustering no es clasificación

- Clustering es aprendizaje no-supervizado
  - Se conocen los datos, no los grupos en que están organizados. El objetivo es encontrar la organización.

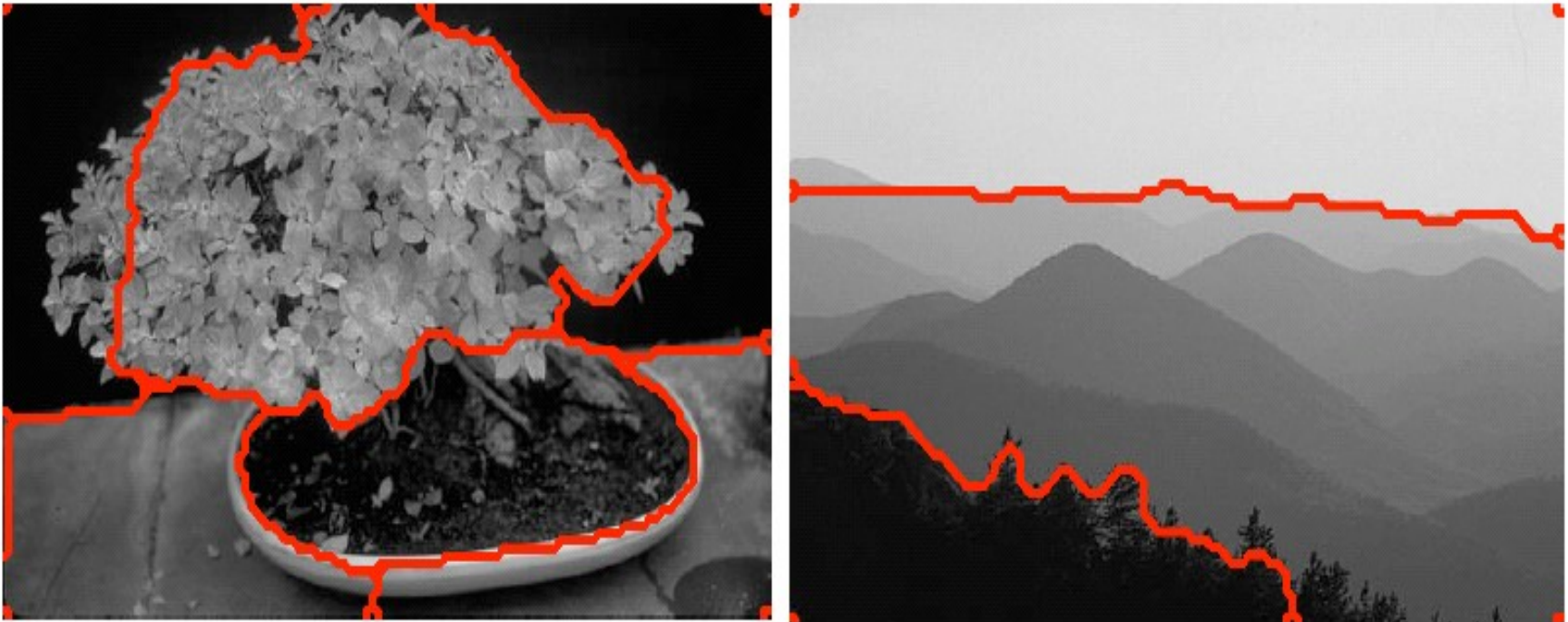




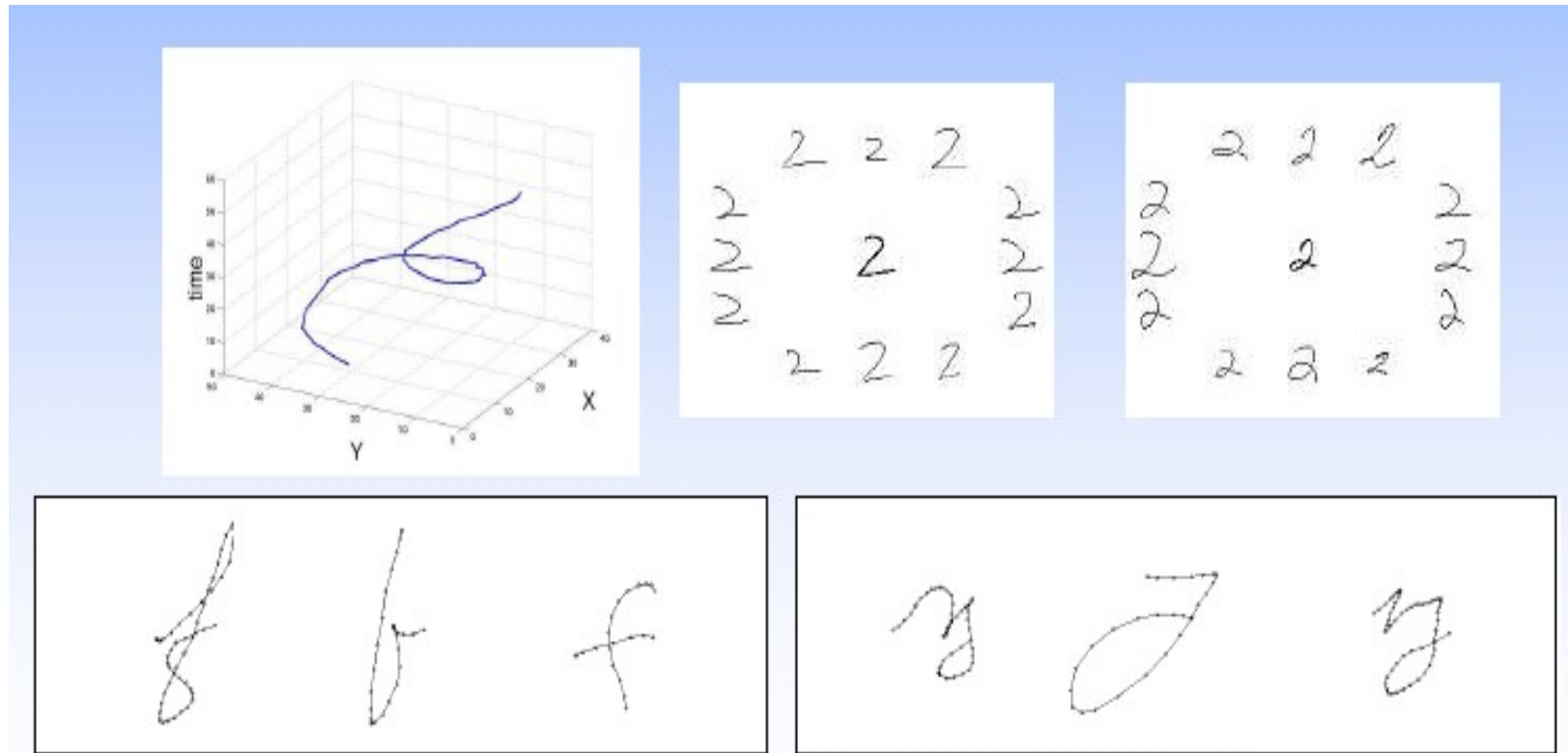
# Ejemplo: Expresión de genes



# Ejemplo: Segmentación de imágenes

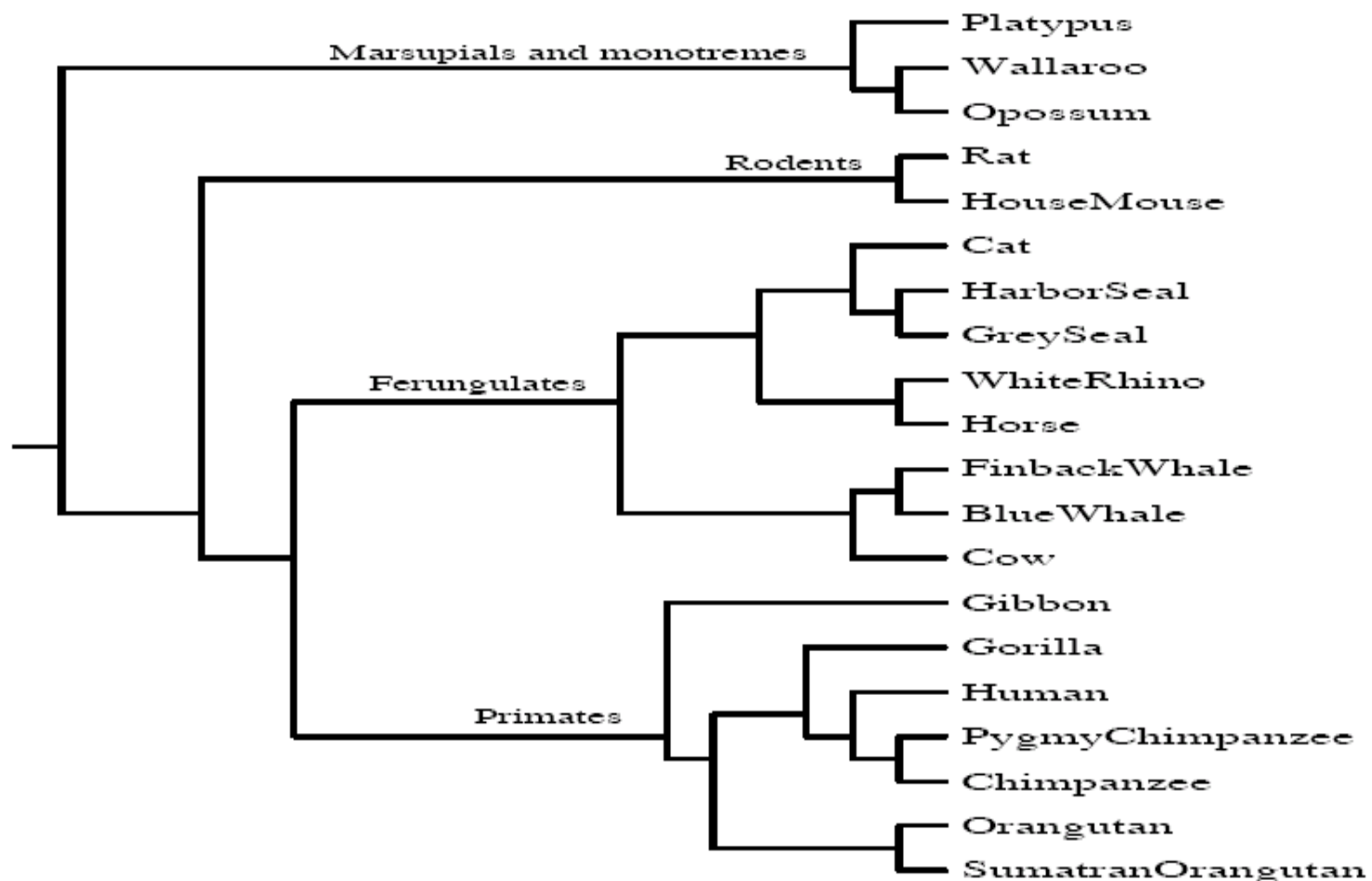


# Ejemplo: Identificación de estilos de escritura.



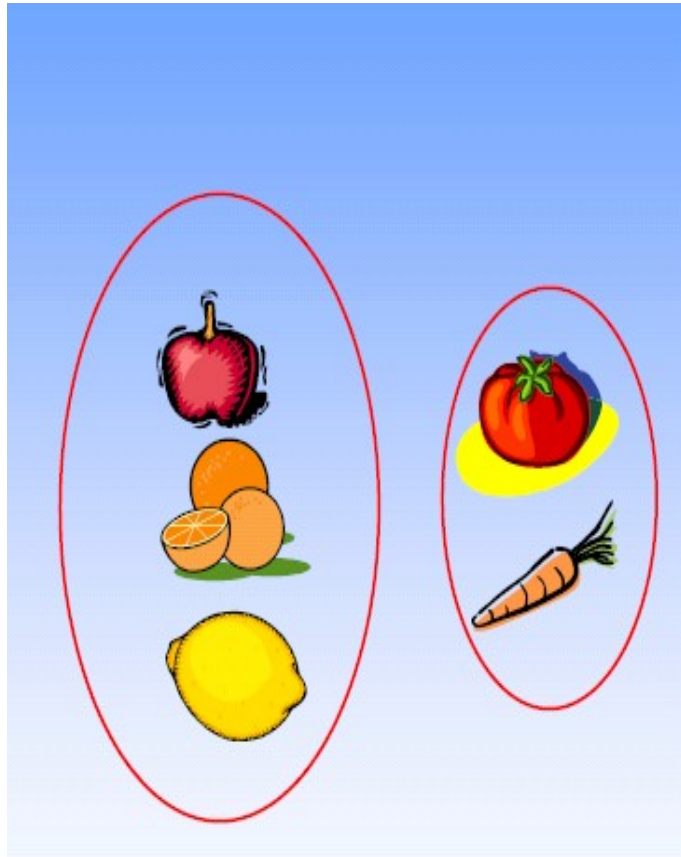
# Ejemplo:

## Distancia genética entre animales

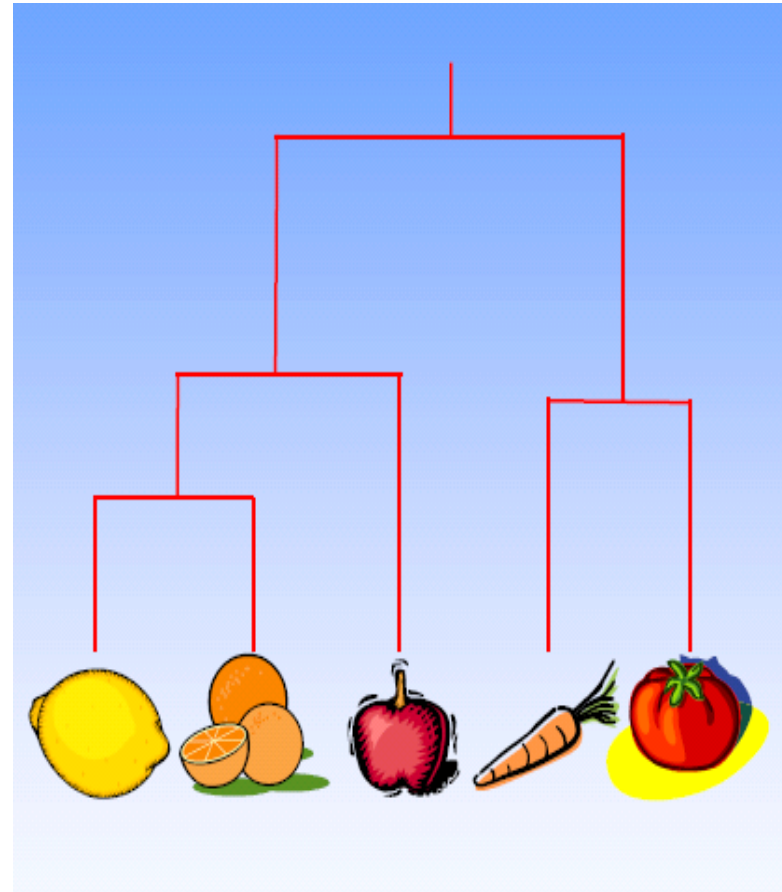




# Dos clases de algoritmos



Divisivos

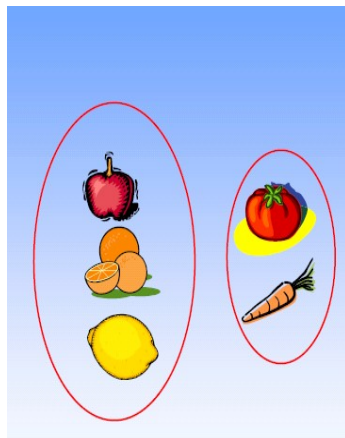


Jerárquicos

# Dos clases de algoritmos

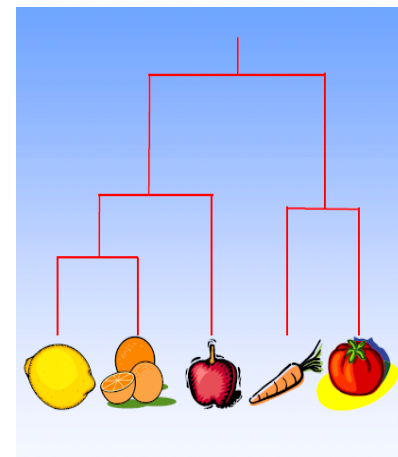
## Divisivos

- “Clustering plano”:  
Clustering como una partición del espacio.
- Queremos la partición “más significativa” en un número fijo de partes.



## Jerárquico

- El objetivo es construir una anidación de particiones, de la que se puede extraer luego una cantidad dada de partes.



# Desarrollo histórico

- Cluster analysis: En nombre aparece en el título de un artículo de análisis de datos antropológicos (*JSTOR*, 1954).
- Hierarchical Clustering: *Sneath (1957), Sorensen (1957)*
- K-Means: Descubierta independientemente por *Steinhaus (1956), Lloyd (1957), Cox (1957), Ball & Hall (1967), McQueen (1967)*
- Mixture models (*Wolfe, 1970*)
- Métodos de teoría de grafos (*Zahn, 1971*)
- K Nearest neighbors (*Jarvis & Patrick, 1973*)
- Fuzzy clustering (*Bezdek, 1973*)
- Self Organizing Map (*Kohonen, 1982*)
- Vector Quantization (*Gersho and Gray, 1992*)

# Datos para clustering

- Datos vectoriales
  - Dos modos: filas y columnas

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- Matriz de distancias
  - Un modo
- R usa los dos

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$



# Métricas

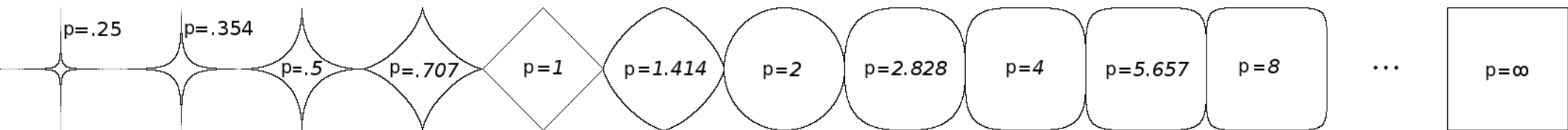
- Para datos vectoriales:

- Minkowski:

$$d(i,j) = \sqrt[p]{\sum |x_{id} - x_{jd}|^p}$$

- $p=1$  Manhattan  $d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{iq} - x_{jq}|$

- $p=2$  Euclidean  $d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{iq} - x_{jq}|^2)}$

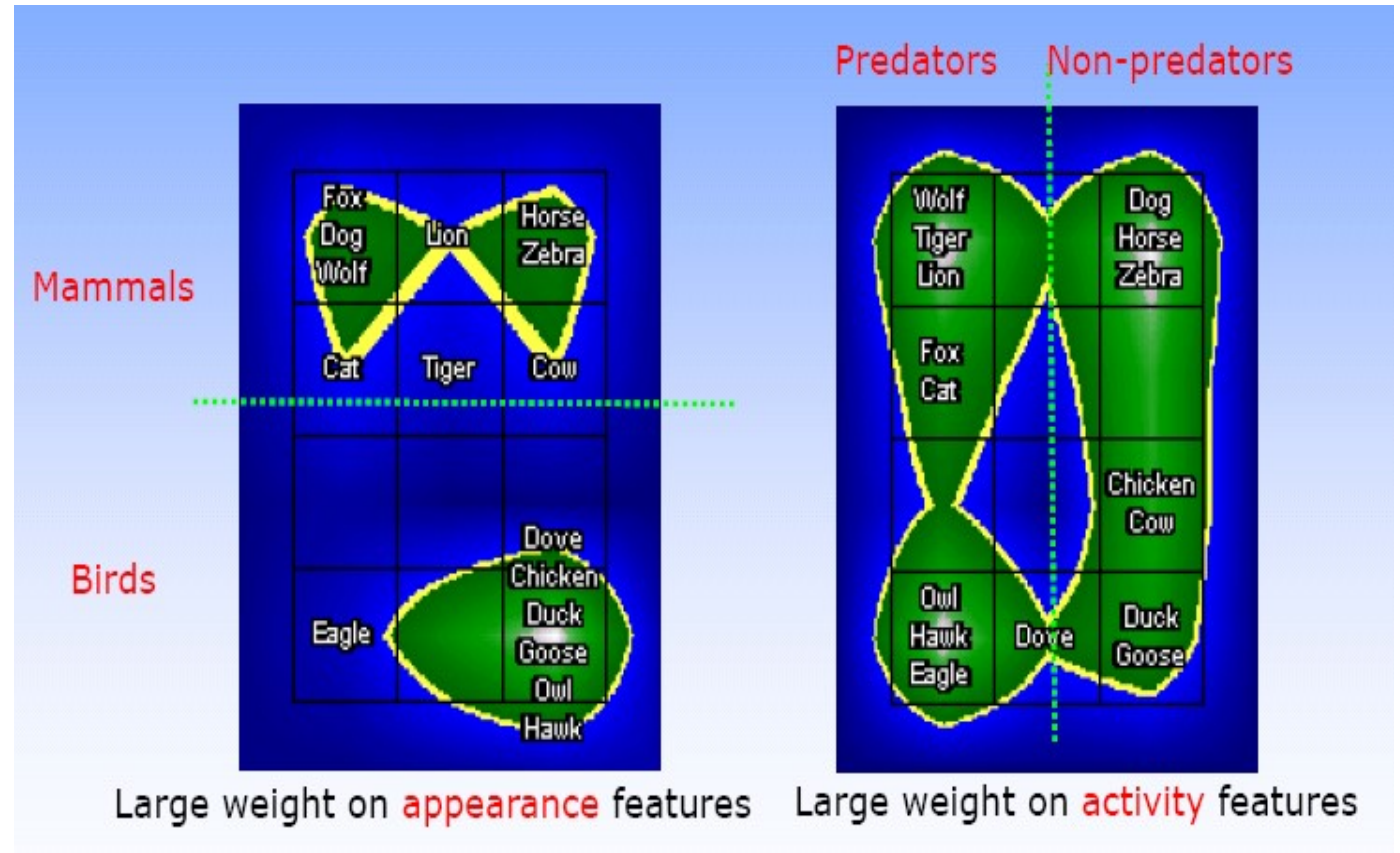


# Métricas (2)

- Para datos vectoriales (otras):
  - Información mutua
  - Correlación
  - Coseno
- Para datos binarios, ordinales o categóricos se definen medidas particulares
- Se pueden definir métricas para tipos especiales
  - Videos, imágenes, texto, etc...

# Pesado de las variables

- 16 animales
- 13 booleanos
- Describen características y comportamiento
- Al cambiar el peso de un grupo de variables a otro cambia totalmente el clustering



# Algoritmo general

- Usar una métrica dada para calcular todas las distancias entre los datos
- Definir una medida de bondad del clustering
  - Por ejemplo, suma de las distancias entre los puntos
- Minimizar la medida de bondad (normalmente con alguna heurística)



# Métodos divisivos



# K-means

- Objetivo: Encontrar una partición de los datos en  $k$  grupos, tal que la distancia media dentro de los puntos de cada grupo sea mínima
  - Grupos apretados, clusters compactos
  - Al minimizar la distancia total dentro de los grupos estamos maximizando la distancia entre los grupos.

# K-means: Planteo

Queremos encontrar una partición tal que:

$$\sum_{j=1}^C \frac{1}{|D_j|^2} \sum_{i \in D_j, l \in D_j} \|X_i - X_l\|^2$$

Se puede ver que es igual a:

$$\sum_{j=1}^C \sum_{i \in D_j} \|X_i - \mu_j\|^2$$

$$\mu_j = \frac{1}{|D_j|} \sum_{l \in D_j} X_l \quad \text{es la media del cluster } j$$

# K-means: Planteo

Queremos el mínimo del costo J:

$$J = \sum_{j=1}^c \sum_{\mathbf{x}_i \in \mathcal{D}_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 = \sum_{j=1}^c \sum_{i=1}^n I(z_i = j) \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$$

Donde los z son las etiquetas de cluster de cada punto.

J es función de los z y los  $\mu$

Si los  $\mu$  están  
fijos y varían los  
z, J es mínimo si:

$$z_i = \arg \min_j \|\mathbf{x}_i - \boldsymbol{\mu}_j\| \quad \forall i$$

Si los  $\mu$  varían, J  
es mínimo si:

$$\frac{\partial}{\partial \boldsymbol{\mu}_j} J = 0 \quad \Rightarrow \quad \boldsymbol{\mu}_j = \frac{\sum_{i=1}^n I(z_i = j) \mathbf{x}_i}{\sum_{i=1}^n I(z_i = j)} = \frac{\sum_{\mathbf{x}_i \in \mathcal{D}_j} \mathbf{x}_i}{|\mathcal{D}_j|}$$



# K-means: Planteo

- Para minimizar  $J$  puedo iterar los dos procesos alternativamente.
- Se puede mostrar que  $J$  desciende siempre.
- Esto se llama minimización alternada
- Si desciende siempre, que garantía tengo???

# K-means: Planteo

- Para minimizar  $J$  puedo iterar los dos procesos alternativamente.
- Se puede mostrar que  $J$  desciende siempre.
- Esto se llama minimización alternada
- Si desciende siempre, que garantía tengo???

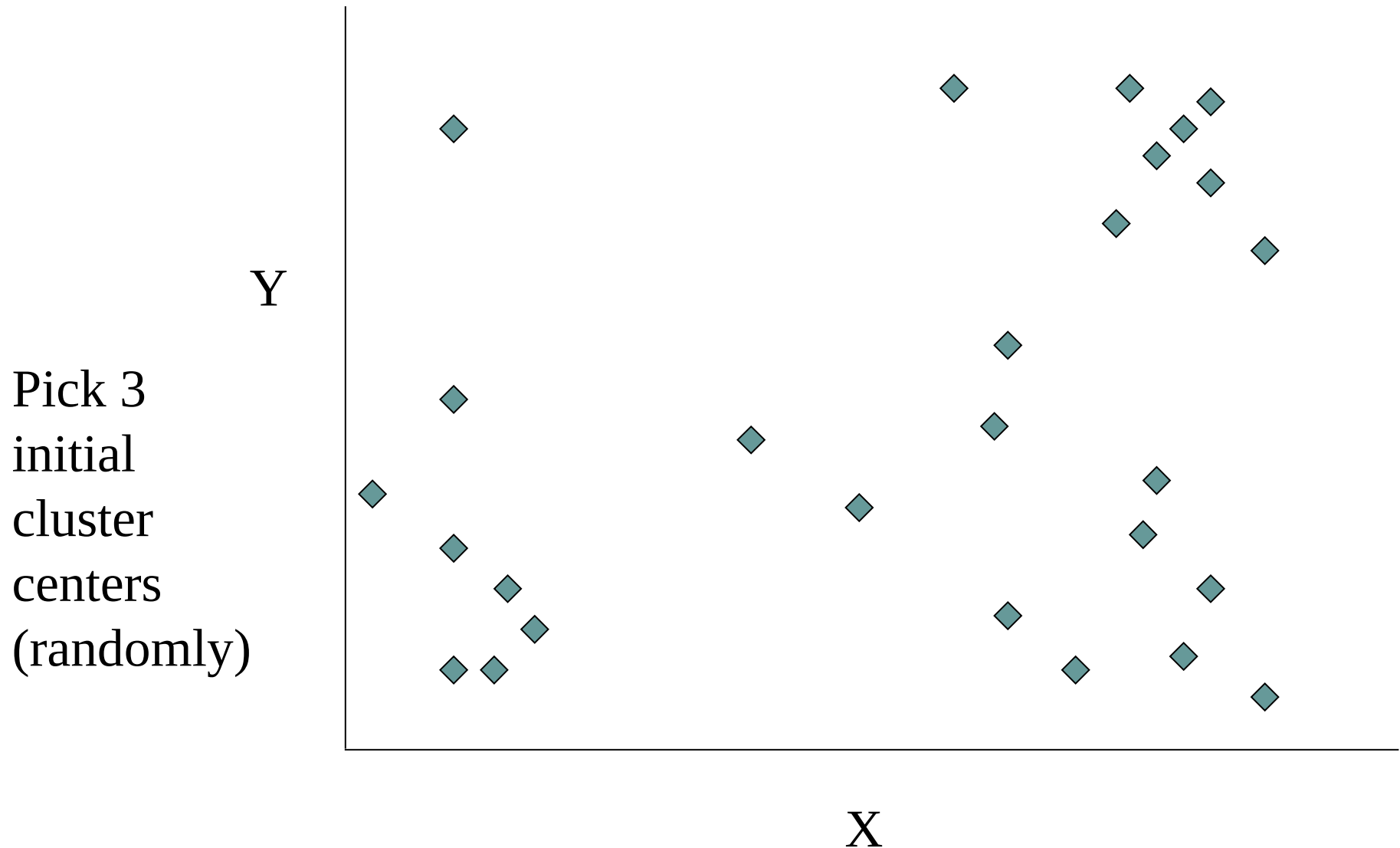
**Voy a encontrar un mínimo LOCAL de  $J$   
en tiempo finito**

# K-means: algoritmo base

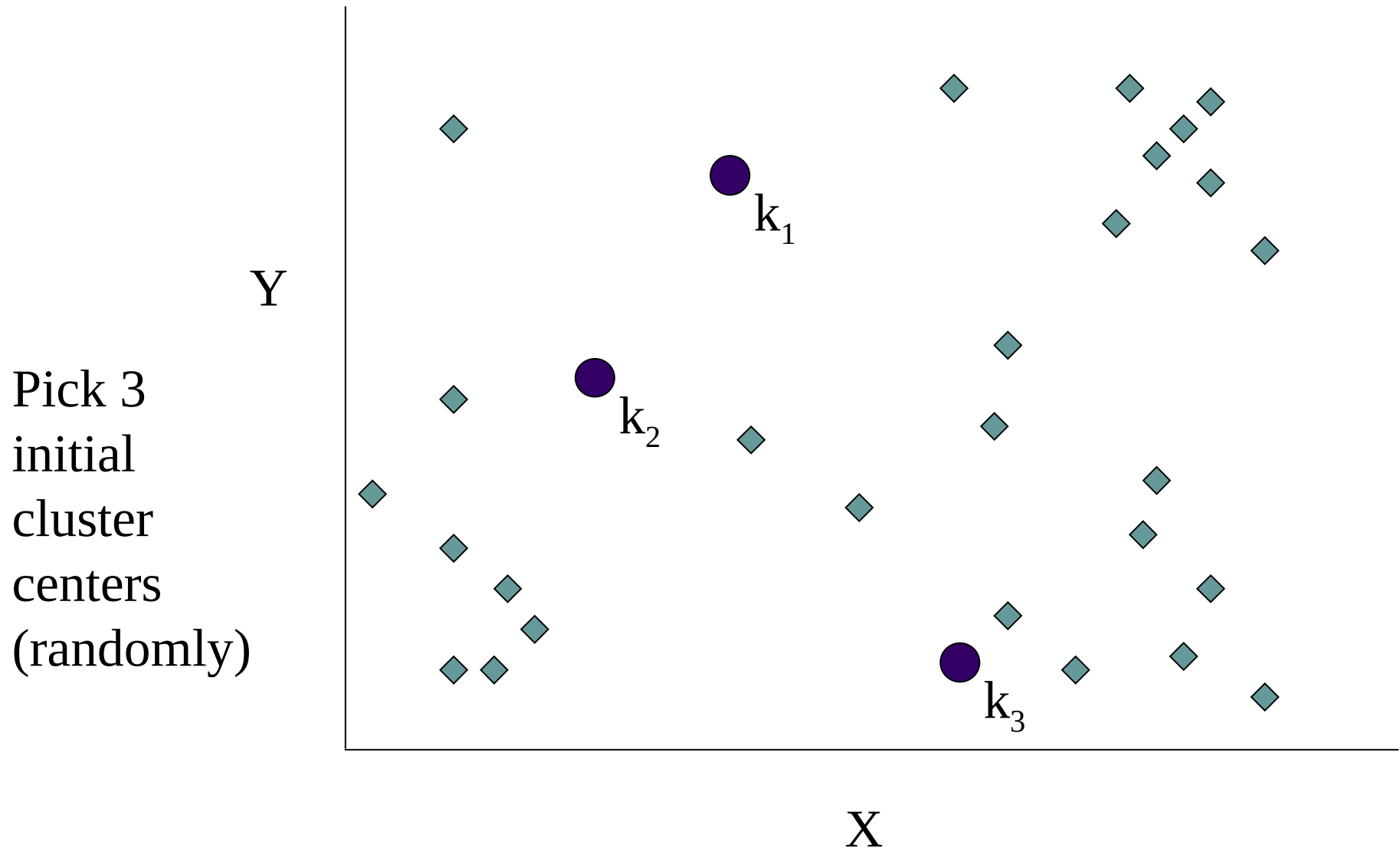
- Empezar con  $k$  centros al azar
- Iterar:
  - Asignar cada punto al centro más cercano
  - Asignar cada centro como la media de sus puntos

Próximas slides: animación del método

# K-means example, step 1

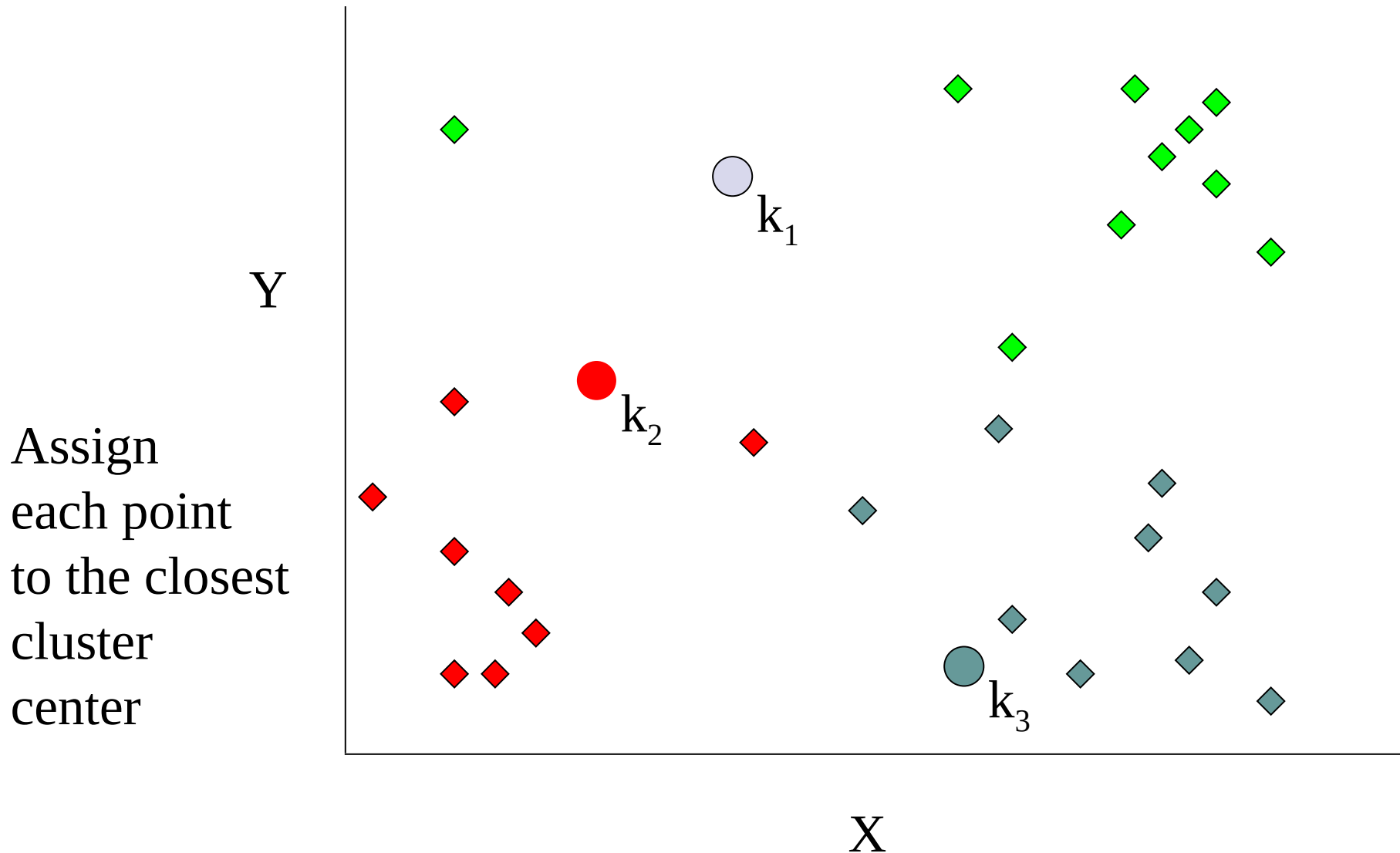


# K-means example, step 1



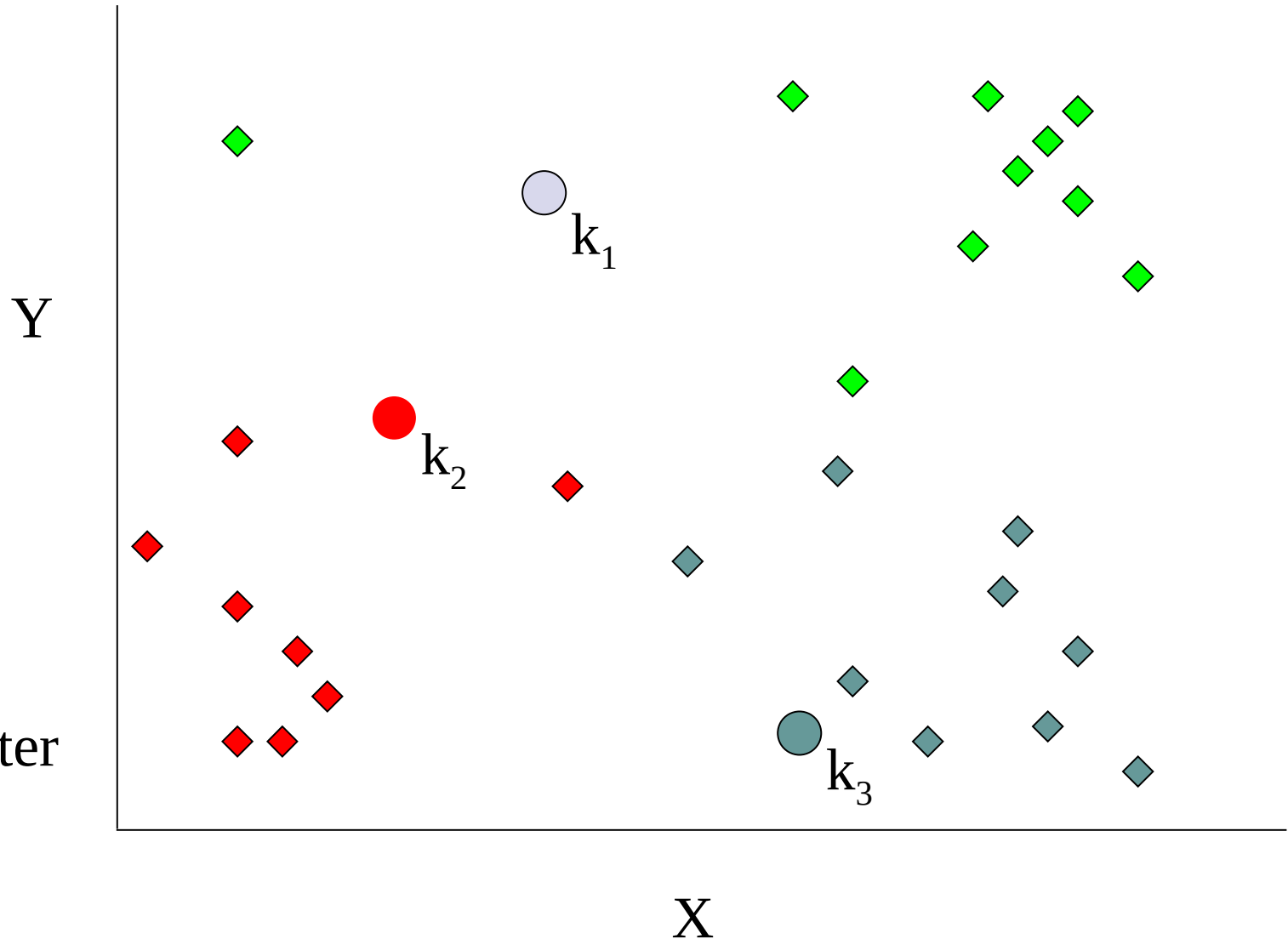


# K-means example, step 2



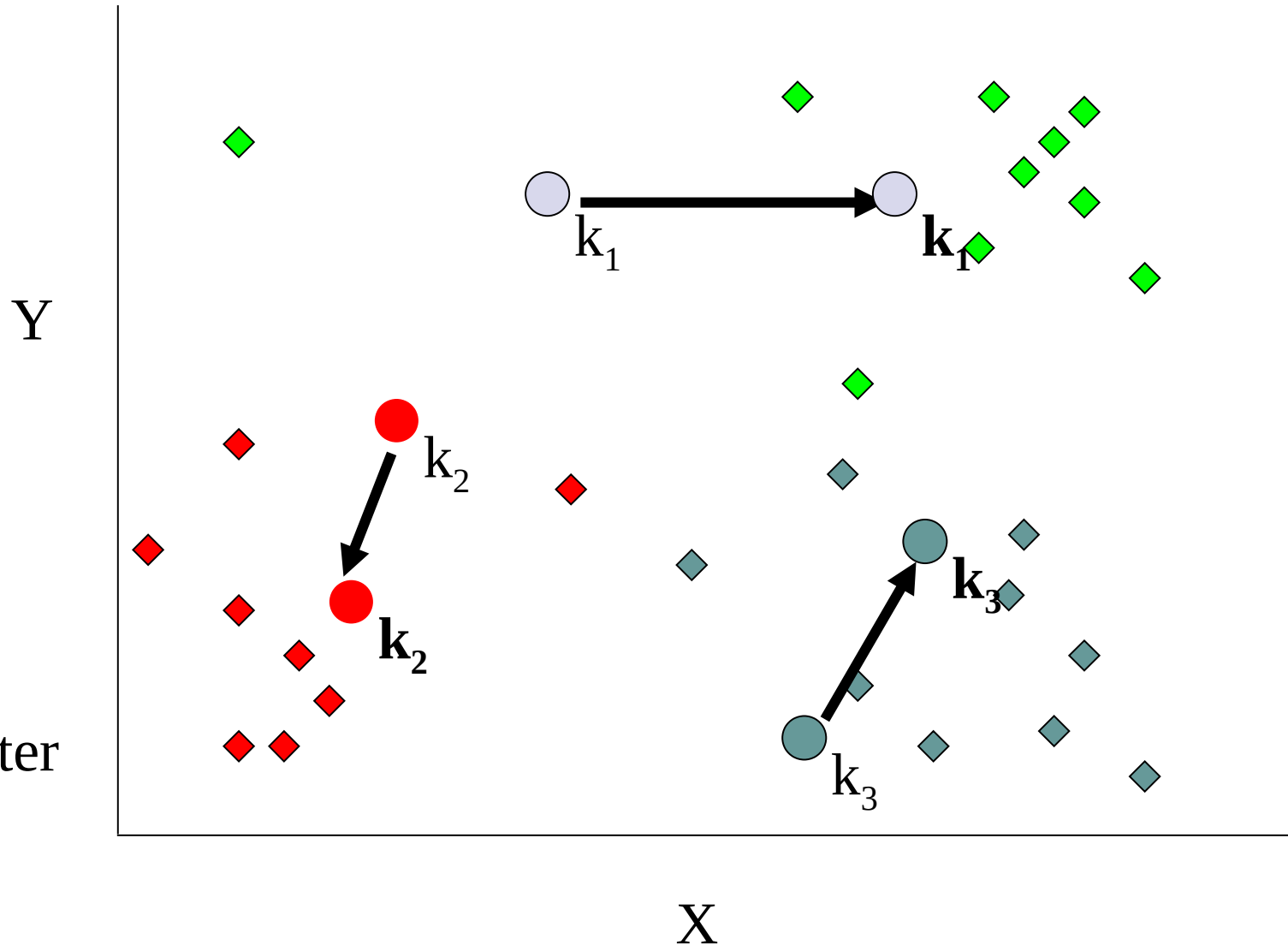
# K-means example, step 3

Move  
each cluster  
center  
to the mean  
of each cluster



# K-means example, step 3

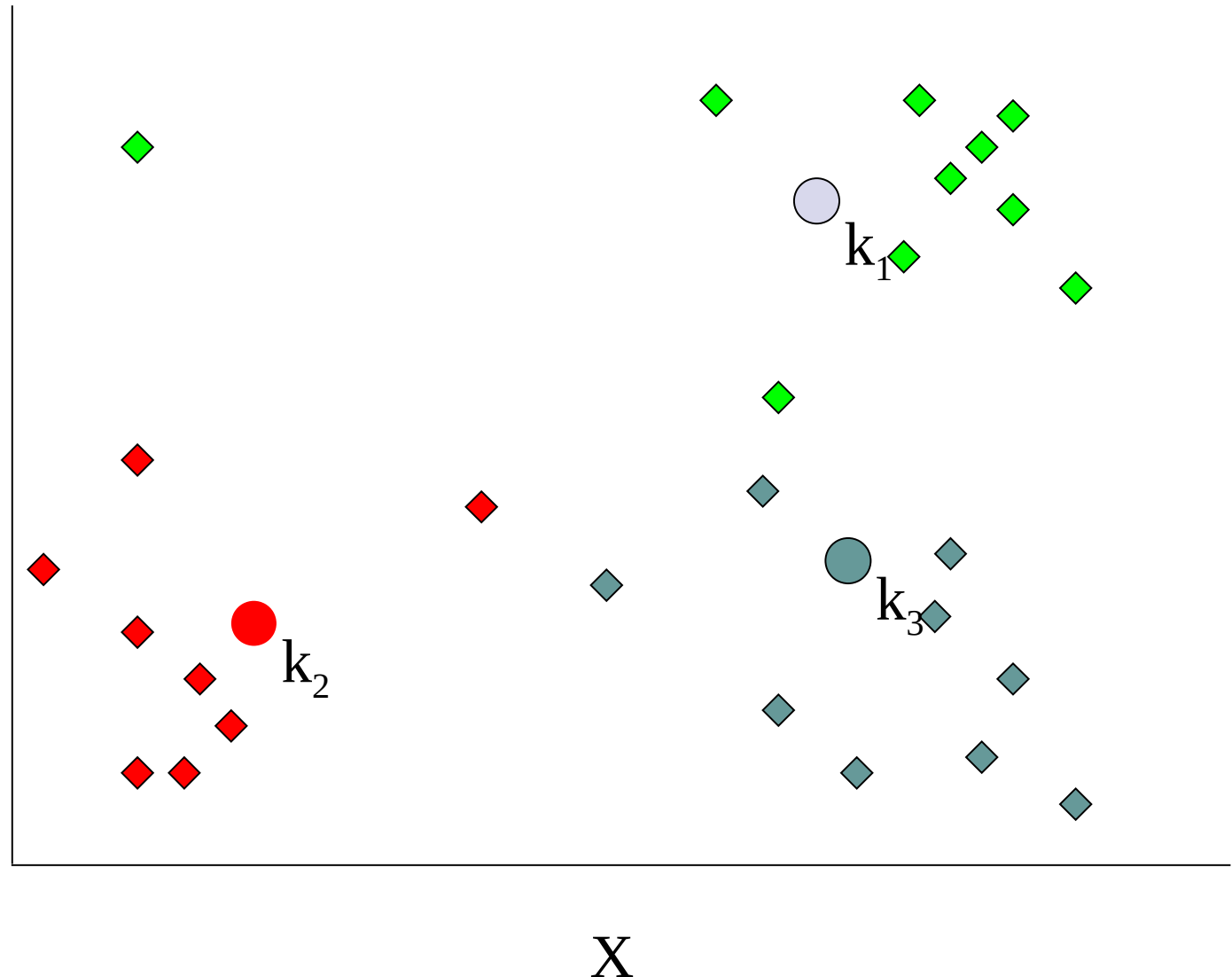
Move  
each cluster  
center  
to the mean  
of each cluster



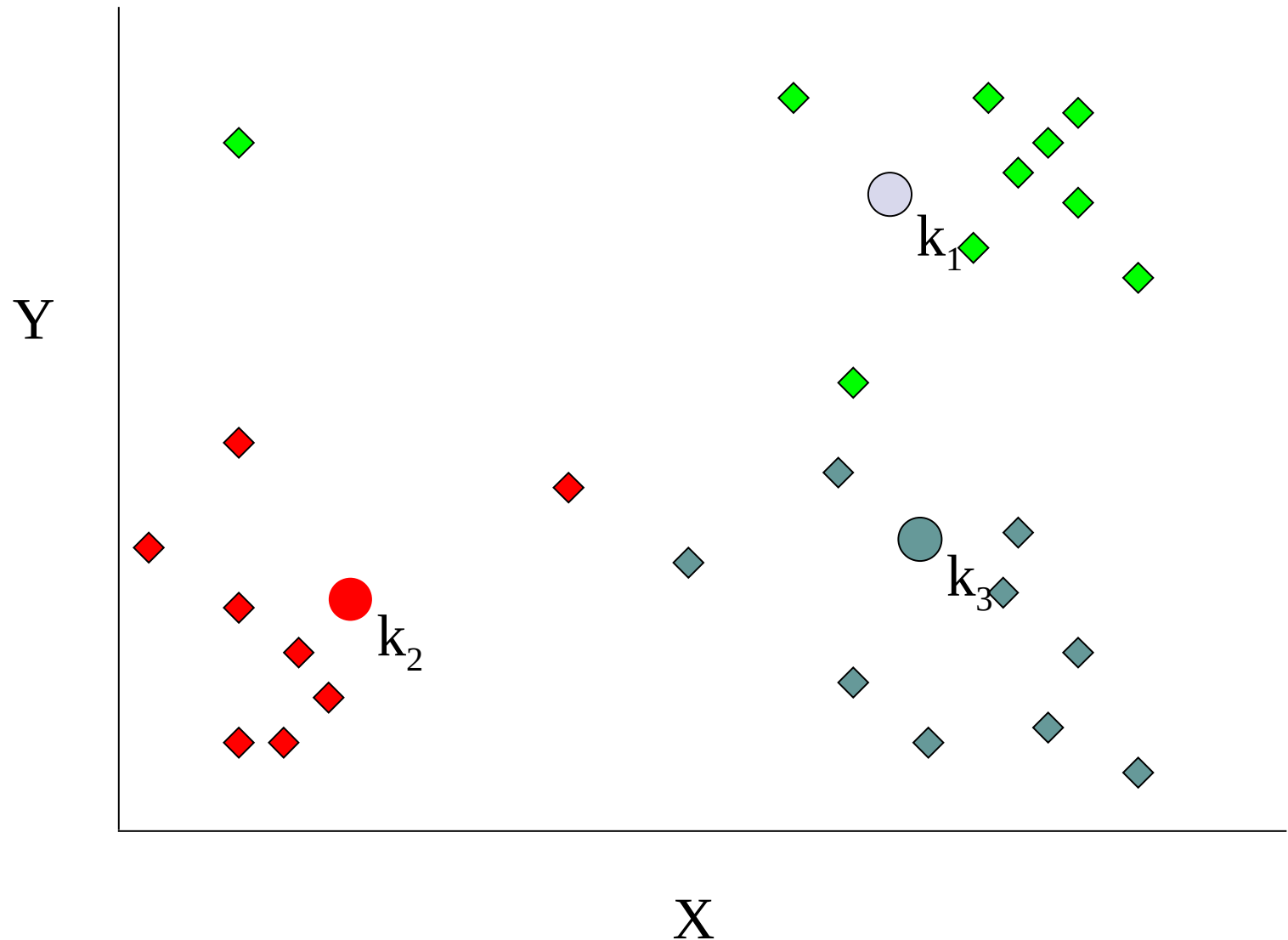
# K-means example, step 4

Reassign  
points  
closest to a  
different new  
cluster center

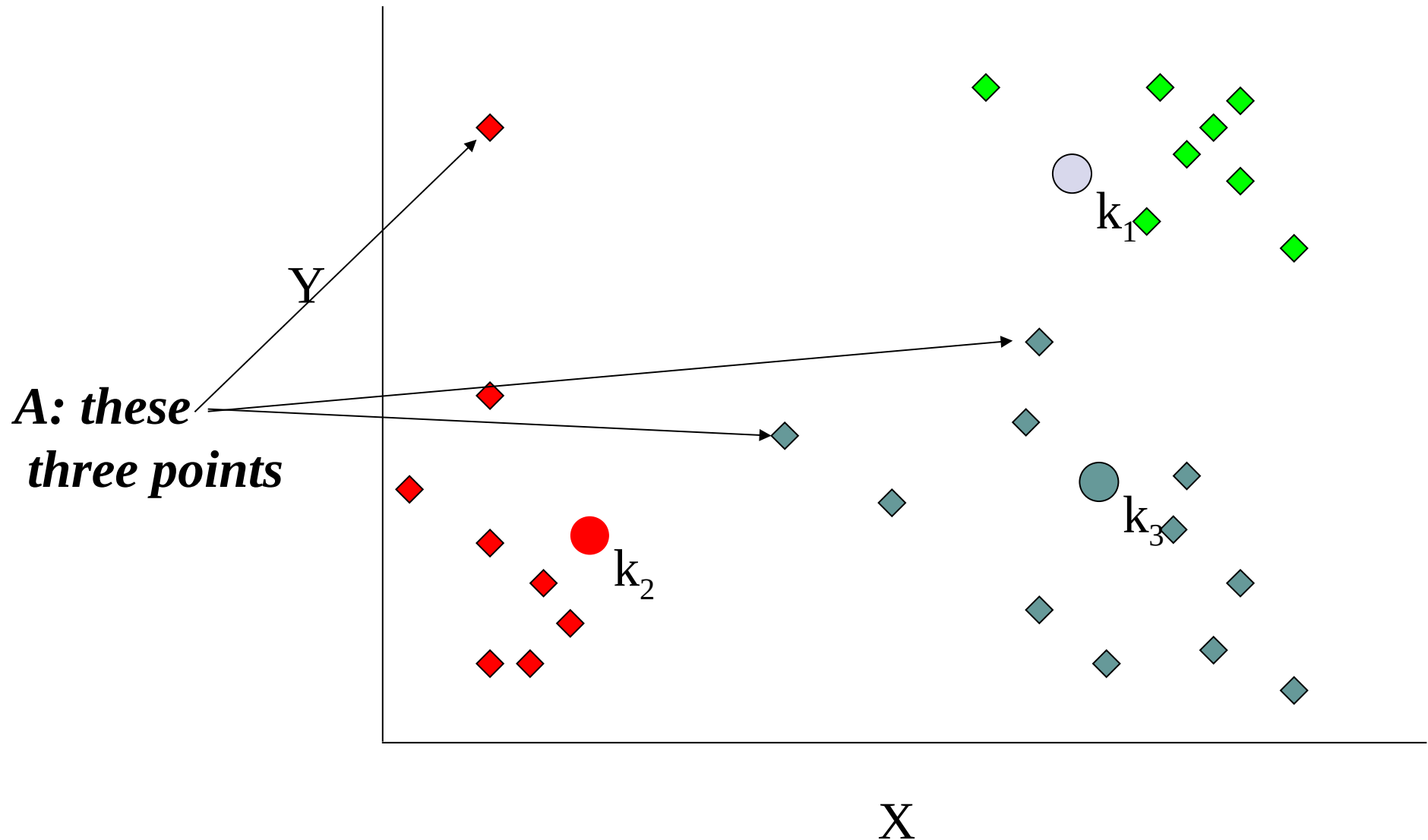
*Q: Which  
points are  
reassigned?*



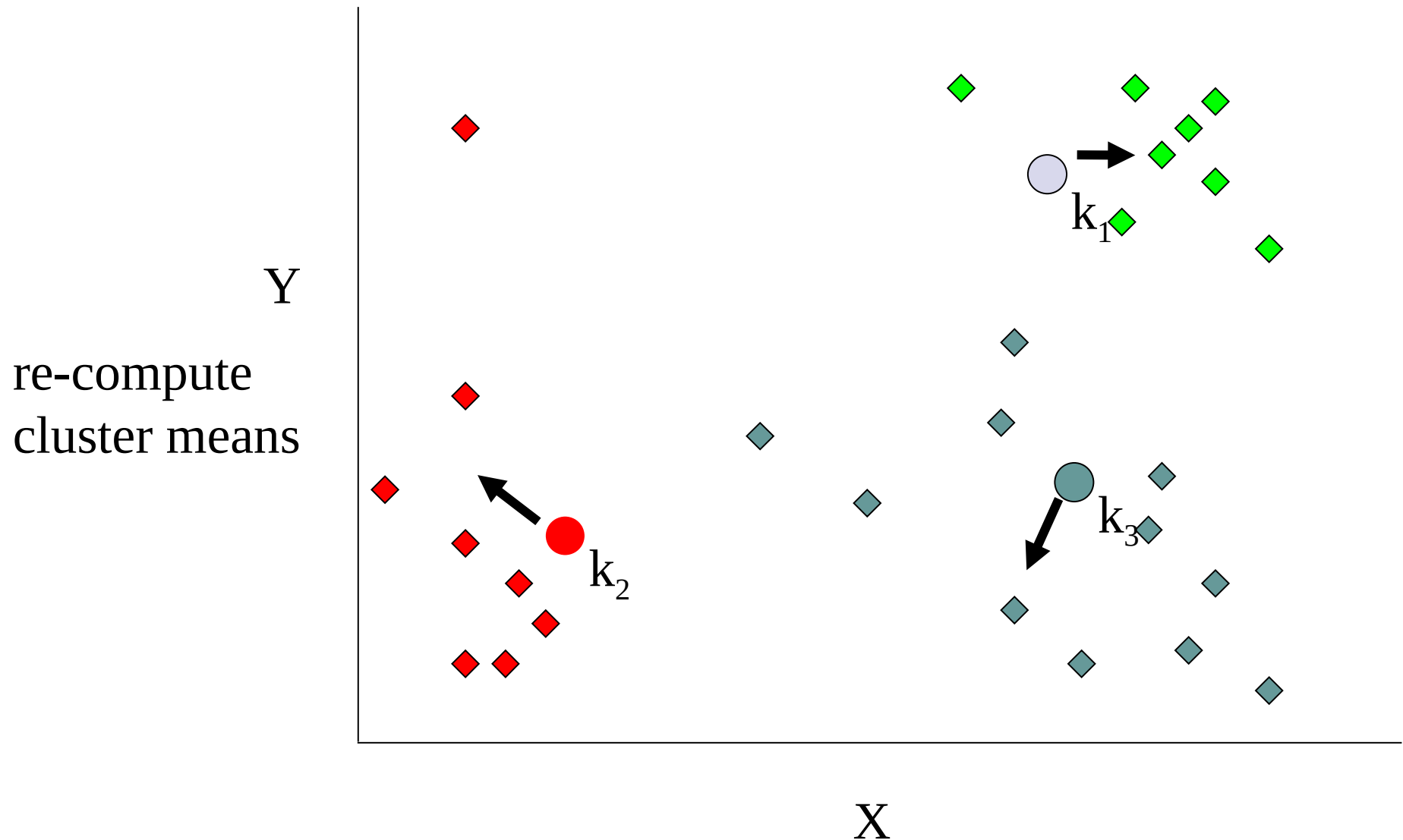
# K-means example, step 4



# K-means example, step 4 ...

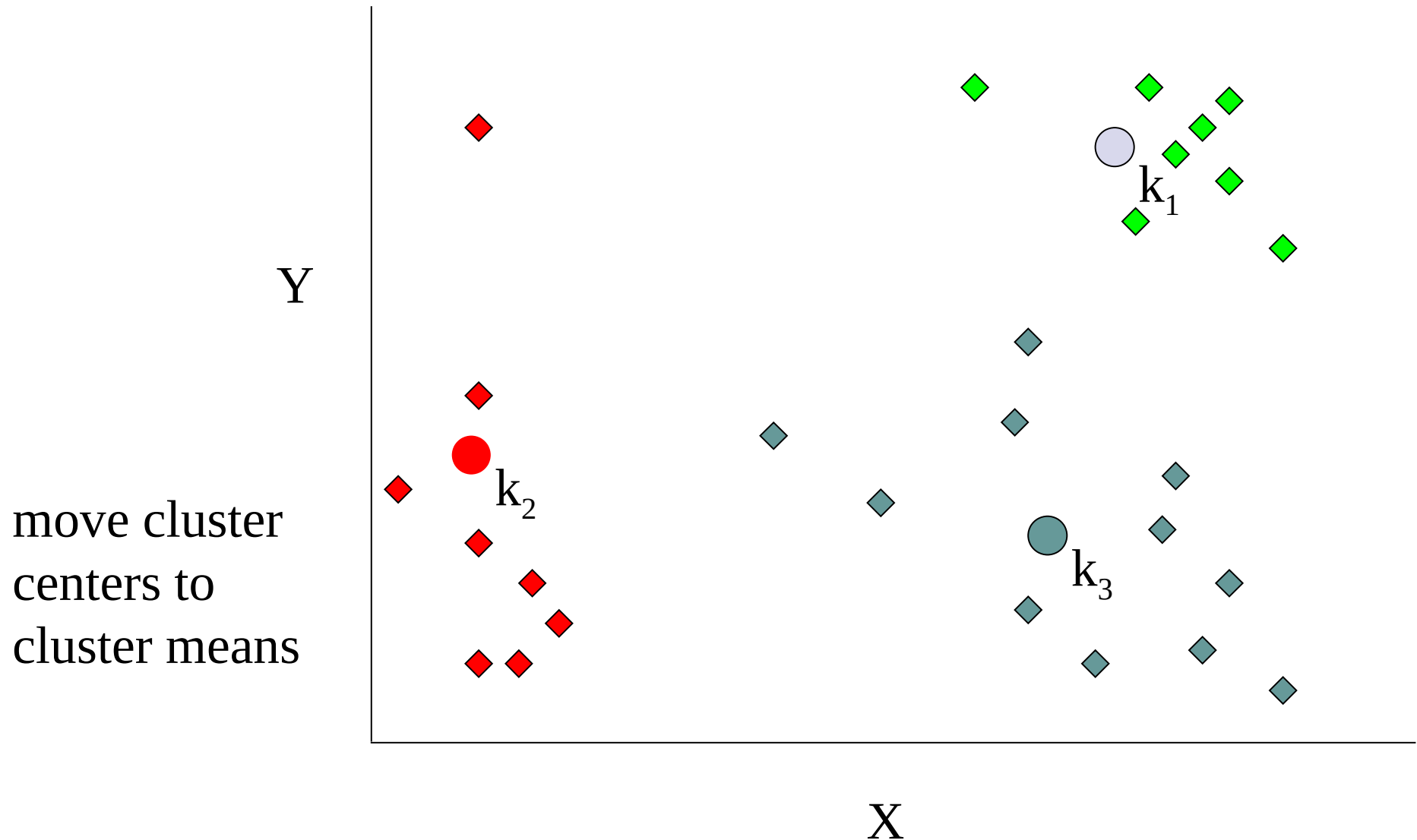


# K-means example, step 4b





# K-means example, step 5

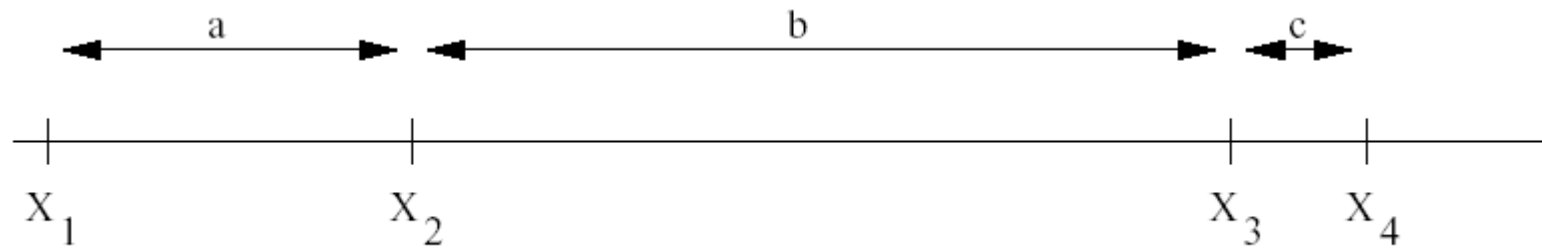


# Fortalezas de k-means

- Eficiente:  $O(tkn)$ , donde
  - $n$  es # objetos
  - $k$  es # clusters
  - $t$  es # iteraciones
  - Normalmente,  $k, t \ll n$
- Garantía de convergencia (a mínimo local)

# Problemas de k-means

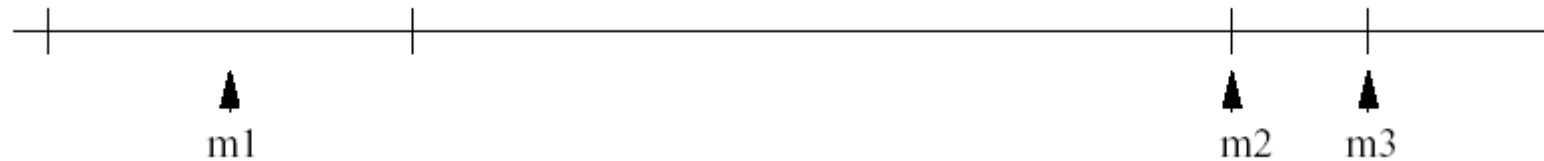
4 puntos en  
1 dimensión,  
3 clusters



Solución óptima:  
 $J=c^2/2$



Solución local:  
 $J=a^2/2$



Cambiando la relación entre  $a$  y  $c$  puede ser tan mala como quiera

# Solución

- Para aumentar la chance de encontrar el mínimo global se usan varias corridas desde distintos valores iniciales, y se compara el  $J$  final
  - Les suena de algún lado?

# Problemas de k-means (2)

- K-means depende fuertemente de los outliers
  - Media de 1, 3, 5, 7, 9 es 5
  - Media de 1, 3, 5, 7, 1009 es 205
  - Mediana de 1, 3, 5, 7, 1009 es 5
  - Ventaja de la Mediana: no la afectan los valores extremos
- K-means solo vale en espacios vectoriales

# Solución (2)

- K-medoids
  - Representar cada cluster por su medoid (es el punto del cluster situado más centralmente)
  - Aplicar la misma iteración que en k-means
  - Soluciona los outliers y vale para espacios arbitrarios
  - PAM (Partitioning Around Medoids, 1987)
  - Mucho más caro computacionalmente  $O(k(n-k)^2)$

# Práctica en R

- Ver archivo de códigos, tiene ejemplos en datos artificiales y reales