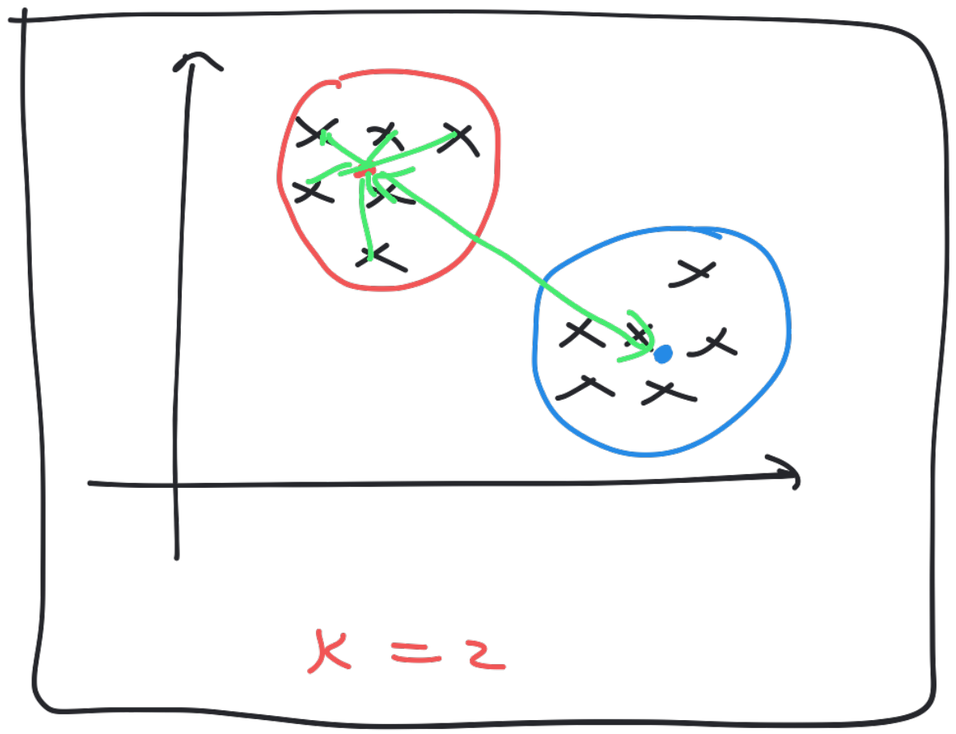
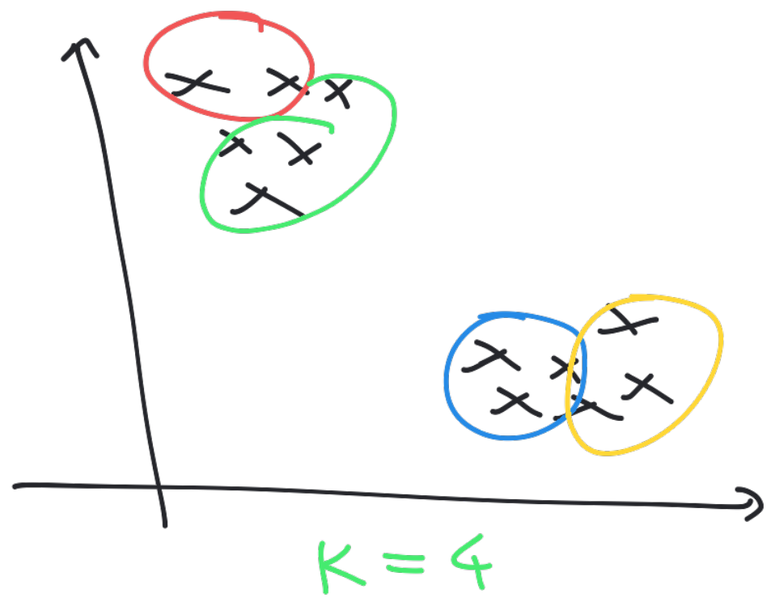


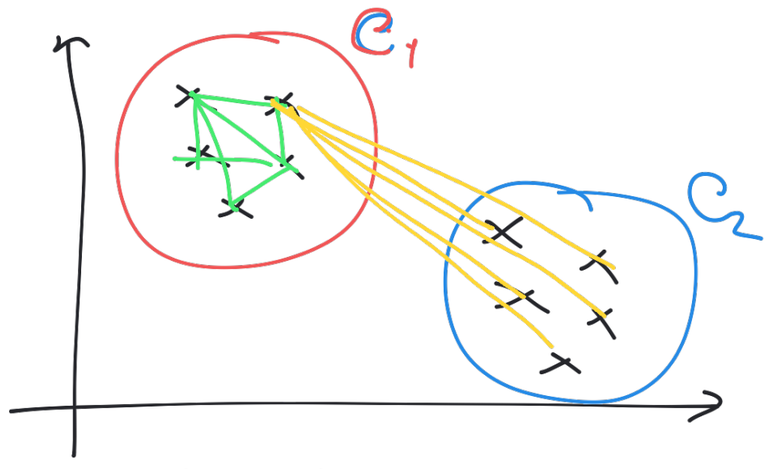

→ we don't know the true labels.

INTERNAL CRITERIA SCHEME:

→ clusters compact

→ distance between cluster centroids is large

## Silhouette index



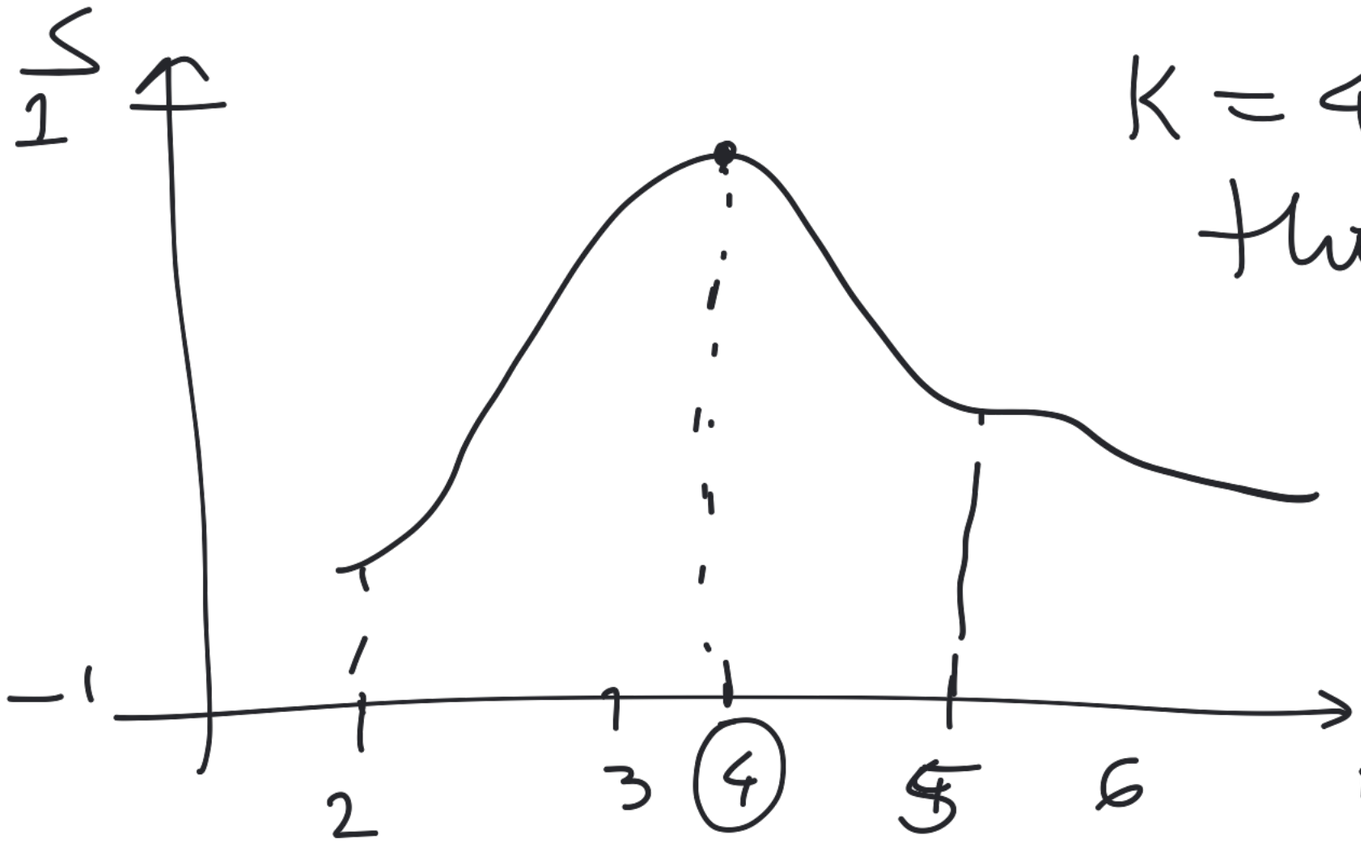

$$X = \{x_i\}_{i=1}^{N}$$

$a_i \equiv$ average distance of point $x_i$ to all the other points of the cluster in which $x_i$ belongs to

$b_i \equiv$ average distance of point $x_i$ to all the points in the other clusters

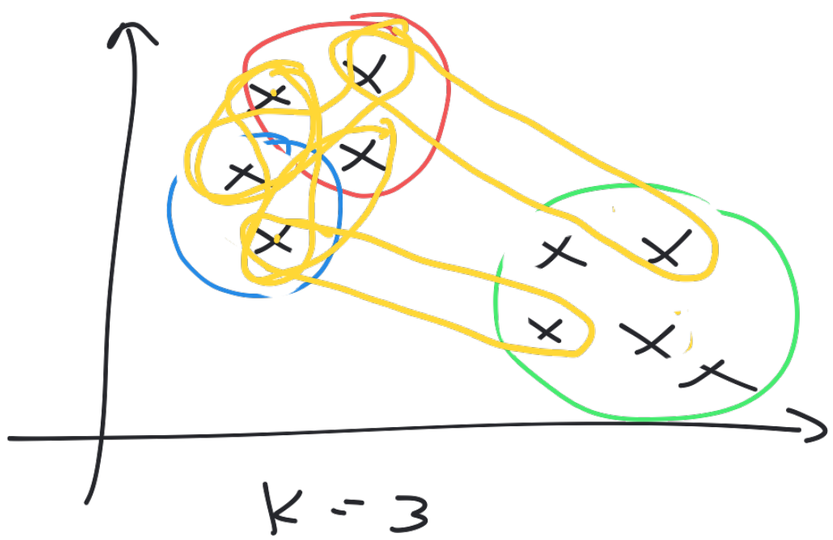
$$-1 \leq S = \frac{1}{N} \sum_{i=1}^{N} \frac{b_i - a_i}{\max(a_i, b_i)} \leq 1$$

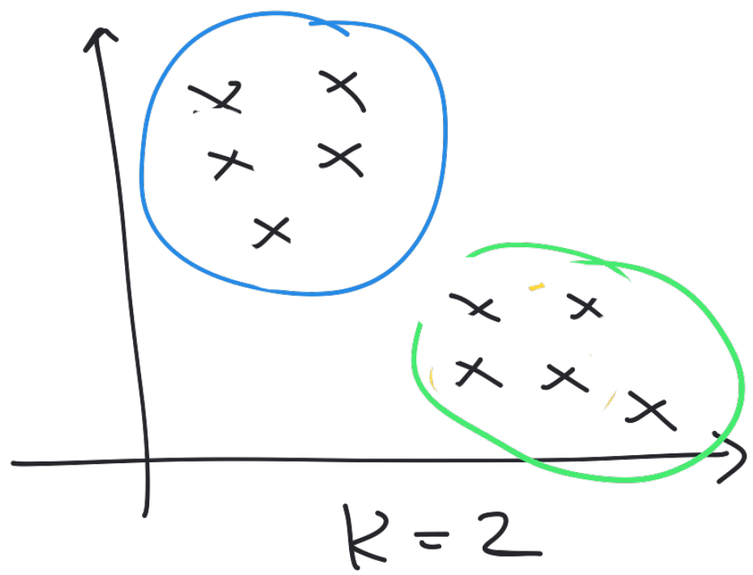
- $S \longrightarrow 1$ returns cluster result with compact clusters far away from each other




K = 4 maximizes the silhouette index, so we choose the clustering results for K=4

# EXTERNAL CRITERIA




clustering result ⟷ TRUE labels

clustering results ⟷ clustering results

SCHEME: — Repeatability
— Consistency.

# RAND INDEX

$$X = \{x_i\}_{i=1}^{N}$$

$$C = \{C_1, C_2, \dots, C_k\}$$

$$D = \{D_1, D_2, \dots, D_s\}$$

$a \equiv$ # of pairs of elements in $X$ that are in the same subset in $C$ and in $D$

↳ previous example: $a = 3$

$b \equiv$ # pairs of elements in $X$ that are in different subset in $C$ and in different subset in $D$

$c \equiv \#$ pairs of elements in X
in the same subset in C
and different subset in D

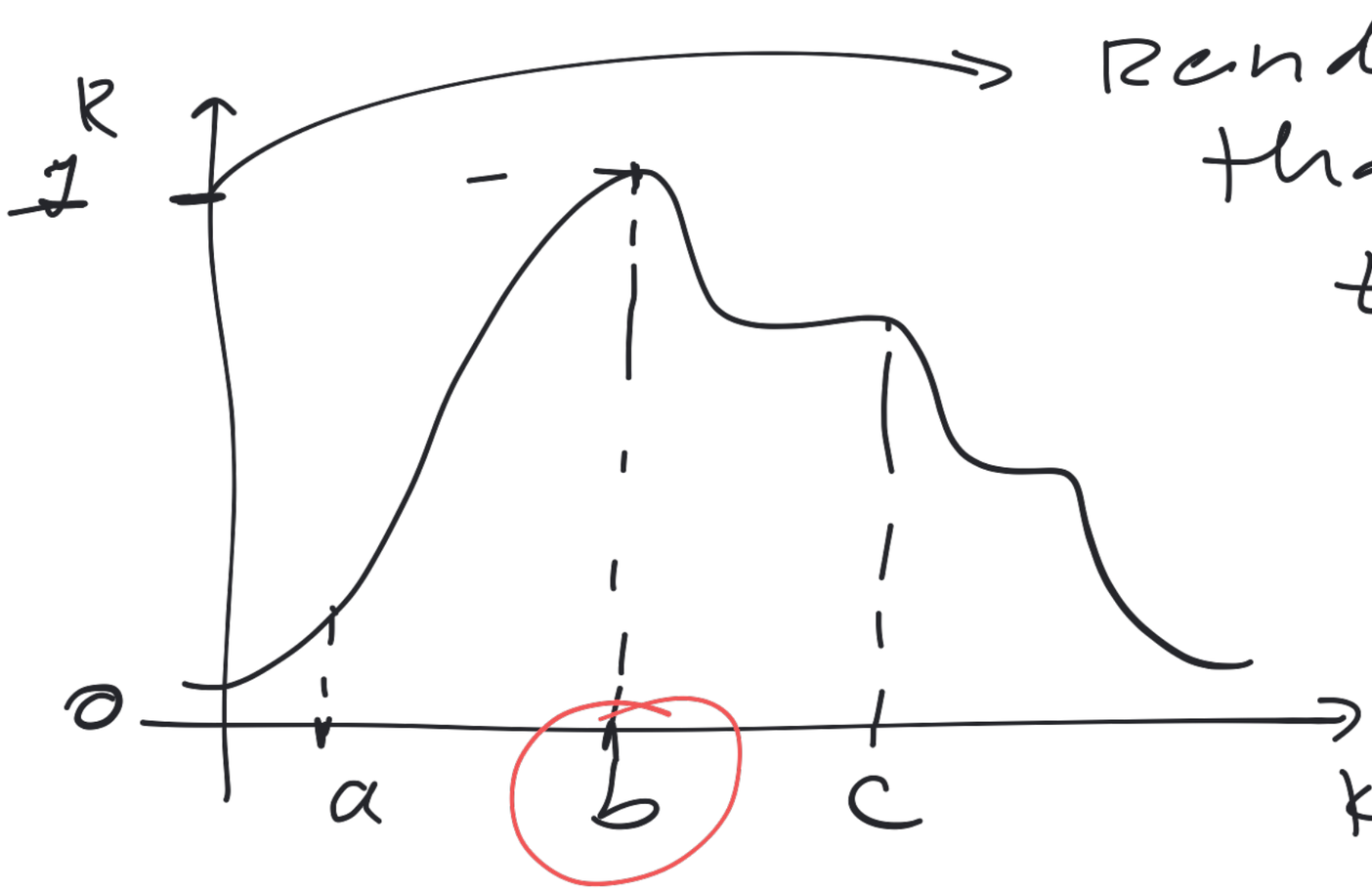
$d \equiv \#$ pairs of elements in X
in different subset in C
but in same subset D

$$0 \leq R = \frac{a+b}{a+b+c+d} \leq 1$$

$a+b \equiv$ "agreement"

$a+b+c+d \equiv$ "every possible pair"

- $R \longrightarrow 1$ we have agreement between clustering results.




Rand index that compares two clustering techniques for any particular k

$k = b$