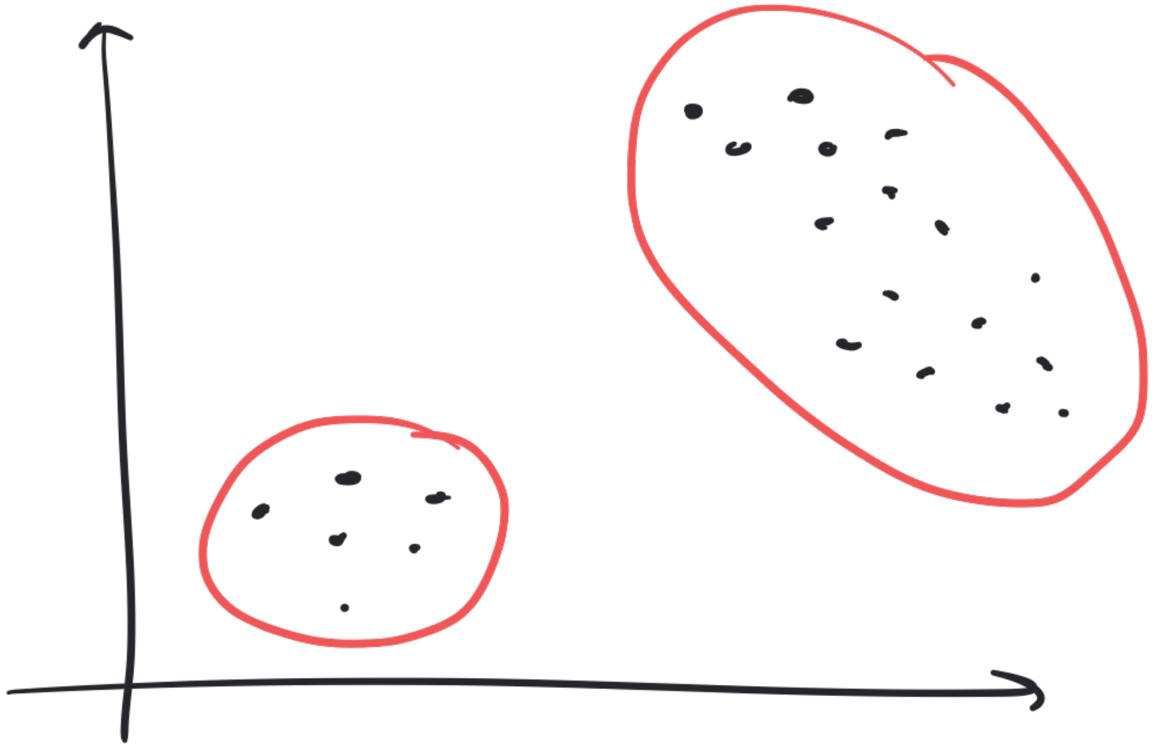
Internal Criteria

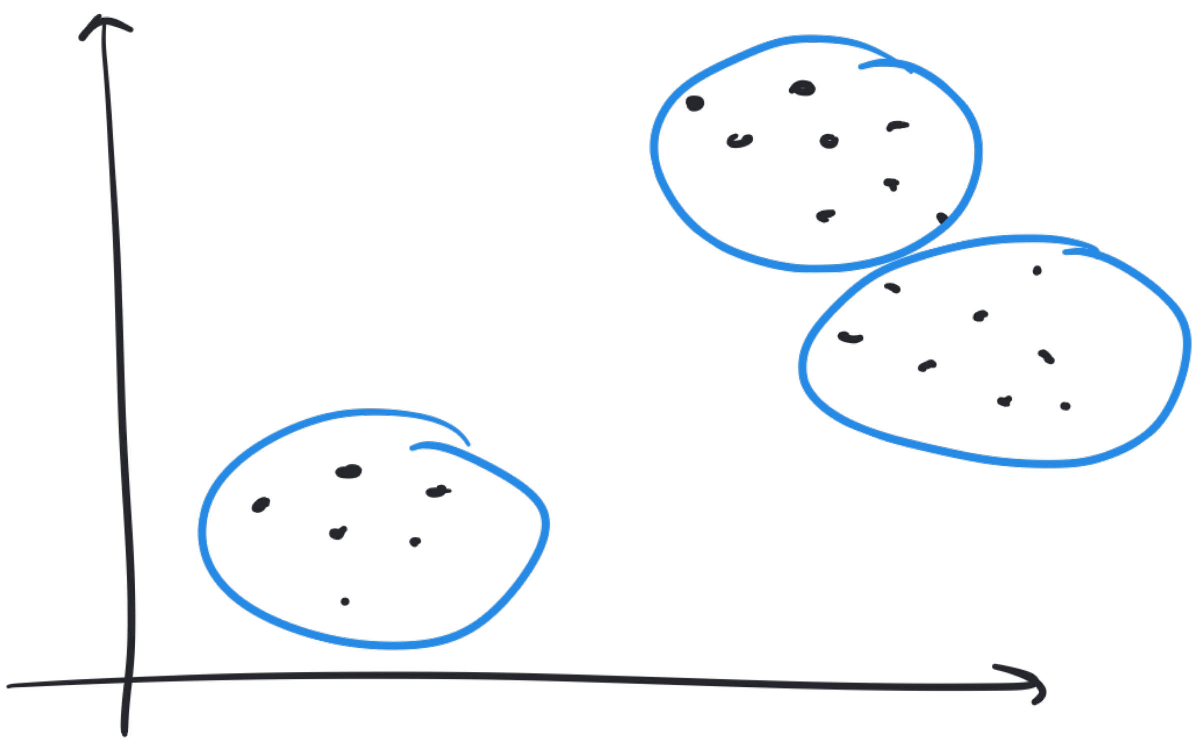
Silhouette Index

- compactness of clusters
- separability between clusters

$K = 2$



$K = 3$



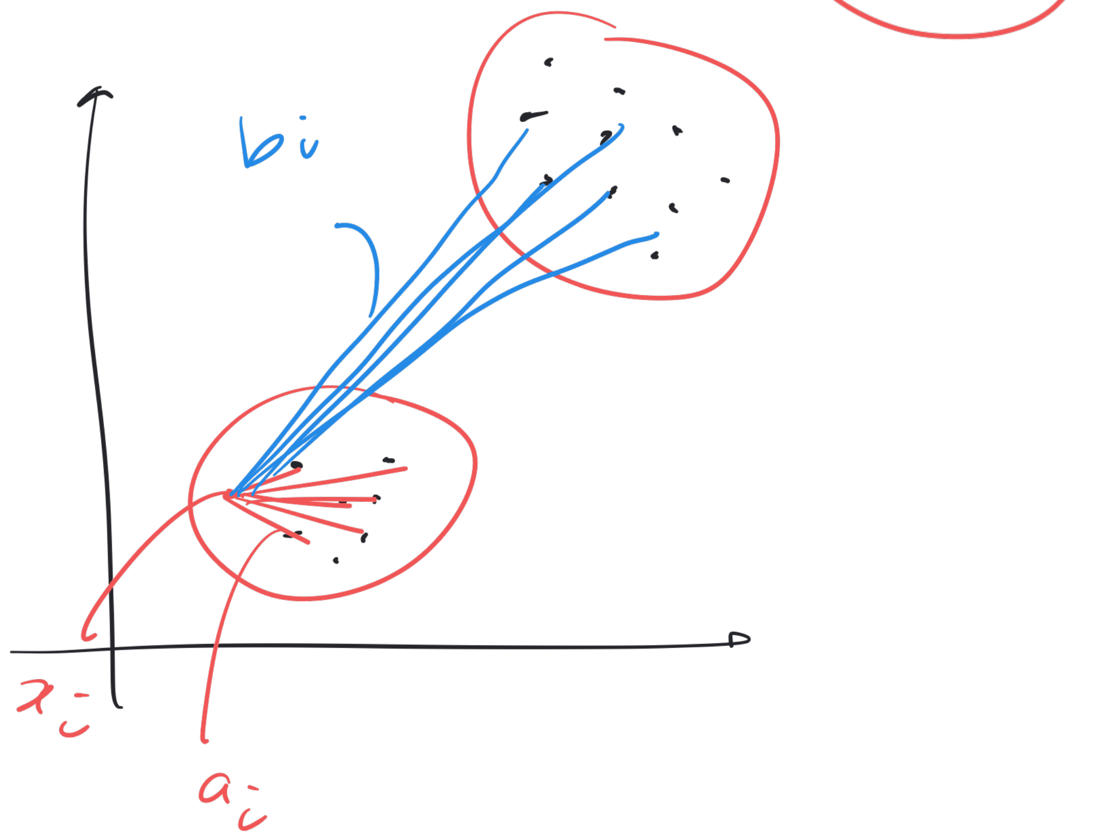
$$s = \frac{1}{N} \sum_{i=1}^{N} \frac{b_i - a_i}{\max(a_i, b_i)}$$

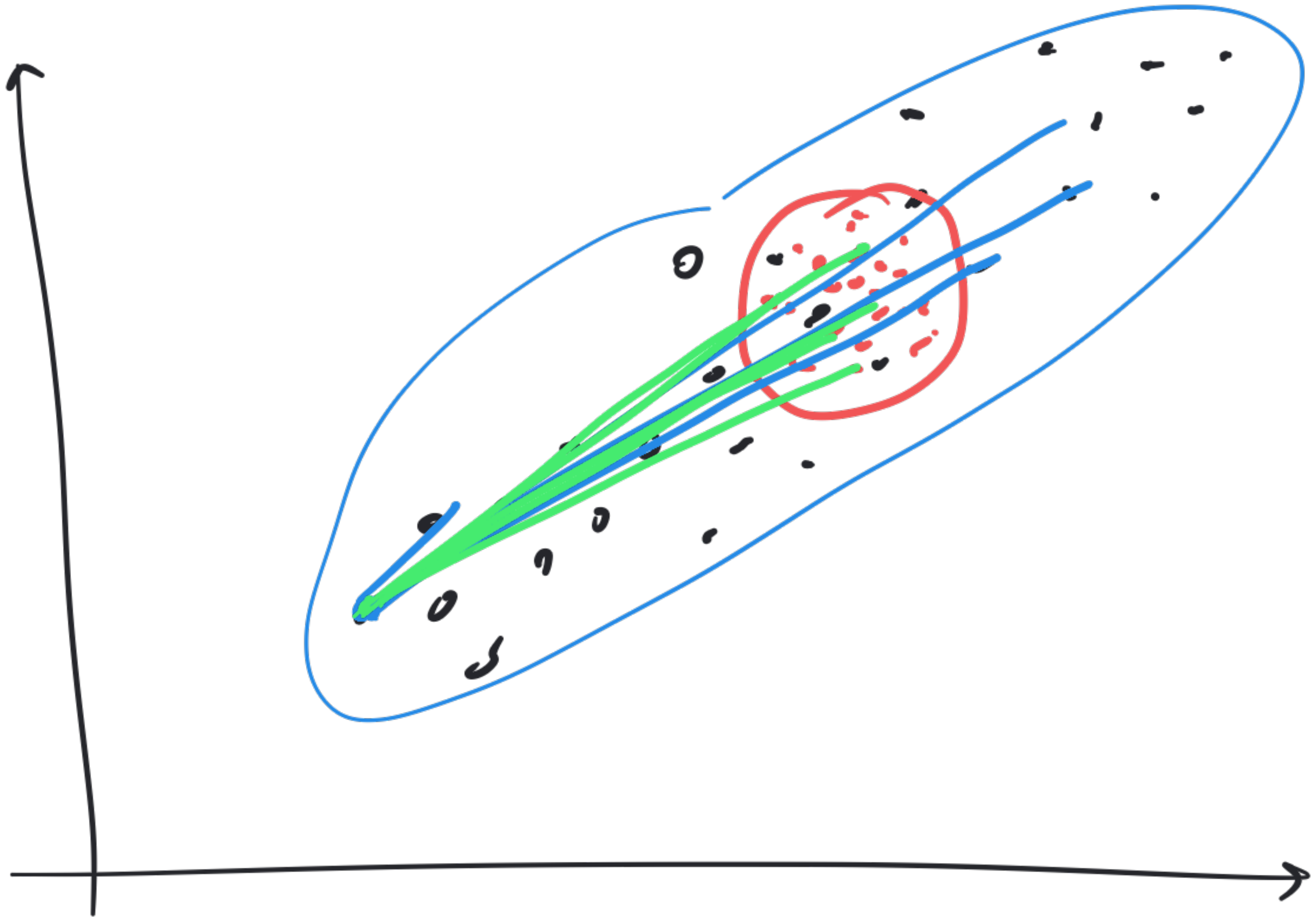
$$\boxed{-1 \le s \le 1}$$

$N \equiv$ # samples

$a_i \equiv$ average distance of point x_i to any other point assigned to the SAME cluster

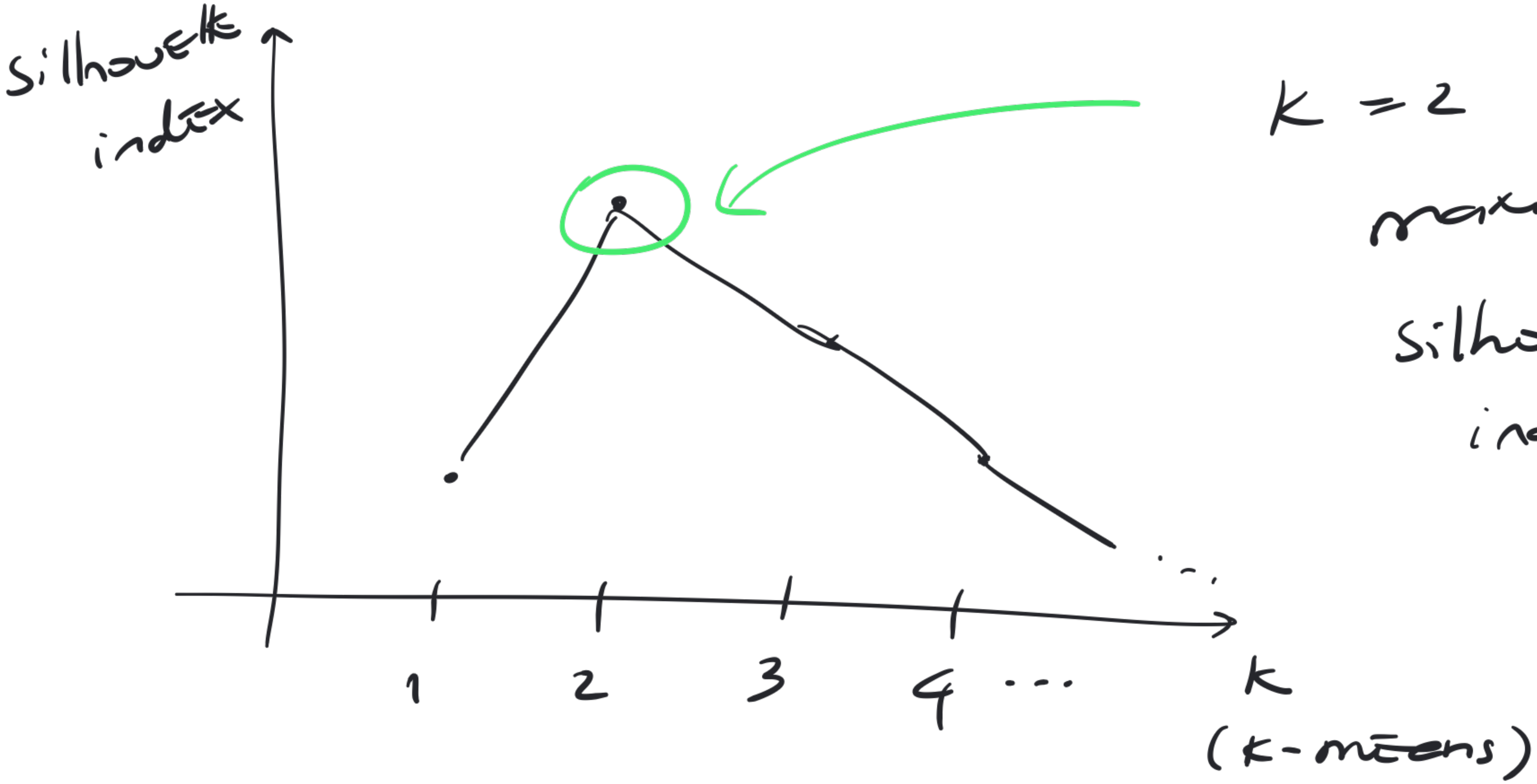
$b_i \equiv$ average distance of point x_i to any other point assigned to a DIFFERENT cluster





Example where s may be

negative.



k = 2 maximizes silhouette index thus it would be selected as best k.

EXTERNAL CRITERIA

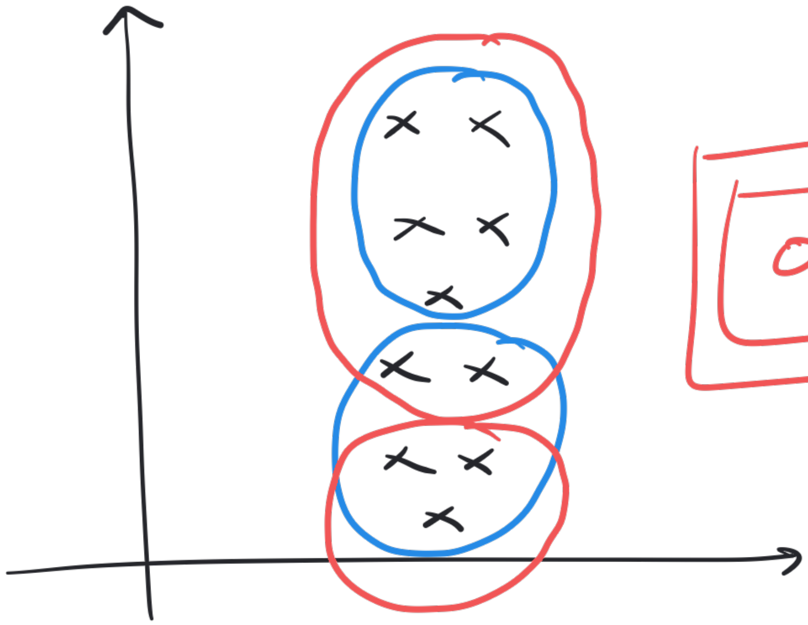
→ compare cluster results from different algorithms.

e.g. K-Means w/ $k = 2$
GMM w/ n_components = 2

→ compare cluster results with ground truth.

RAND INDEX

$$\boxed{R = \frac{a+b}{a+b+c+d}}$$



— RESULT B

— RESULT A, e.g. K-MEANS w/ K = 2

$$\boxed{0 \le R \le 1}$$

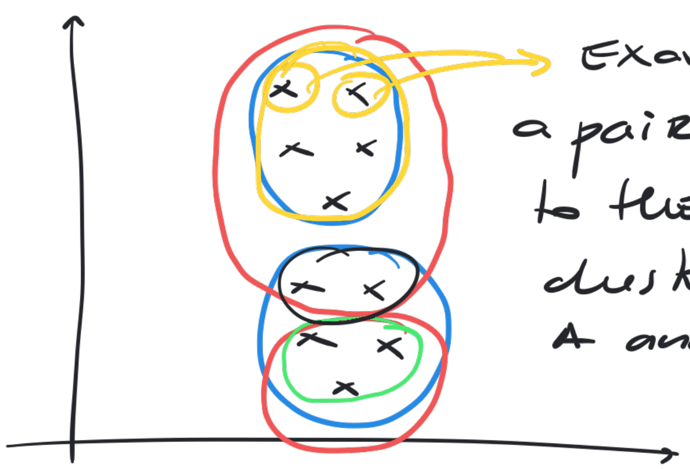
$a \equiv$ # pairs of points assigned to the same cluster in both A and B

$b \equiv$ # pairs of points assigned to a different cluster in both A and B

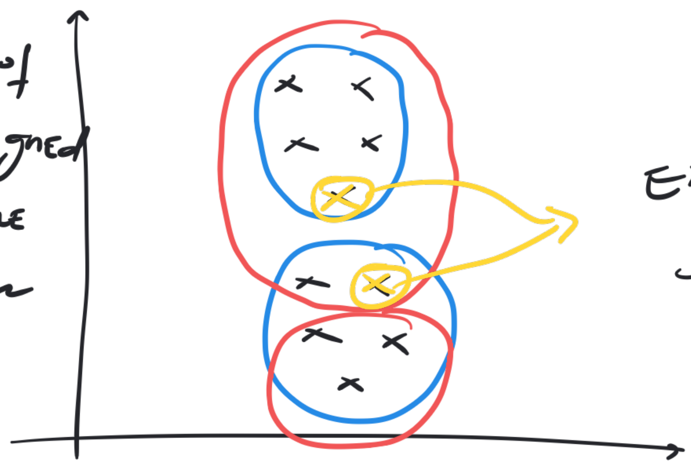
$c \equiv$ # pairs of points assigned to the same cluster in A but different cluster in B.

$d \equiv$ # pairs of points assigned to different cluster in A but the same cluster in B.

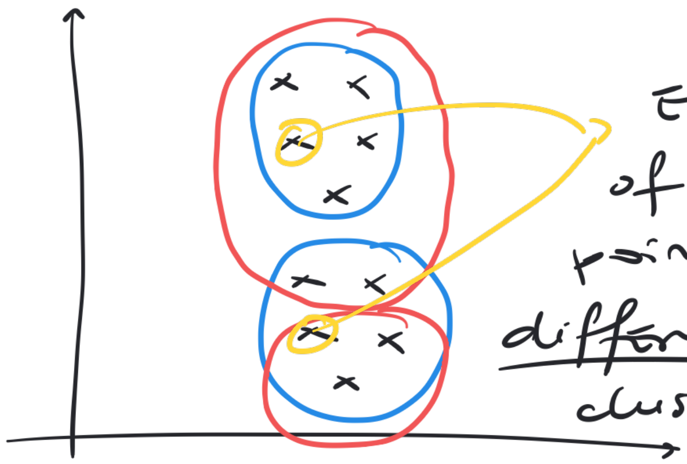
— result A
— result B



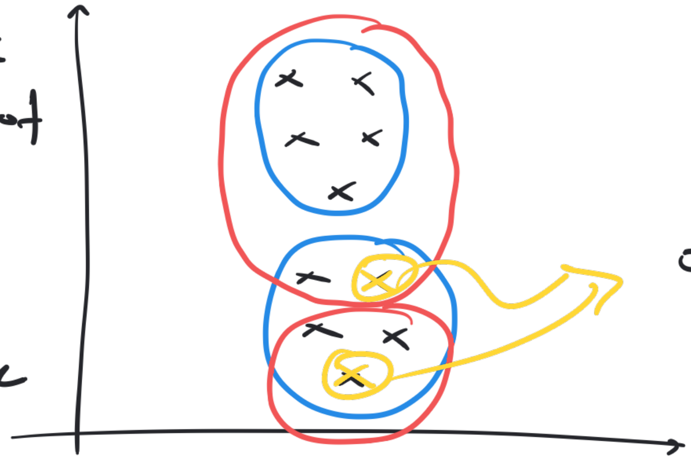
Example of a pair assigned to the same cluster in A and B



Example of pair of points assigned to the same cluster in A but different in B.



Example of pair of points in different clusters in A and B



Example of pair of points assigned to different cluster in A but the same in B.

For example,

$$a = \binom{5}{2} + \binom{3}{2} + \binom{2}{2}$$

$$\binom{5}{2} = C^5_2 = \frac{5!}{2!(5-2)!} \qquad \binom{2}{2} = 1$$

$$a + b \equiv \# \text{agreements}$$

$$c + d \equiv \# \text{disagreements}$$

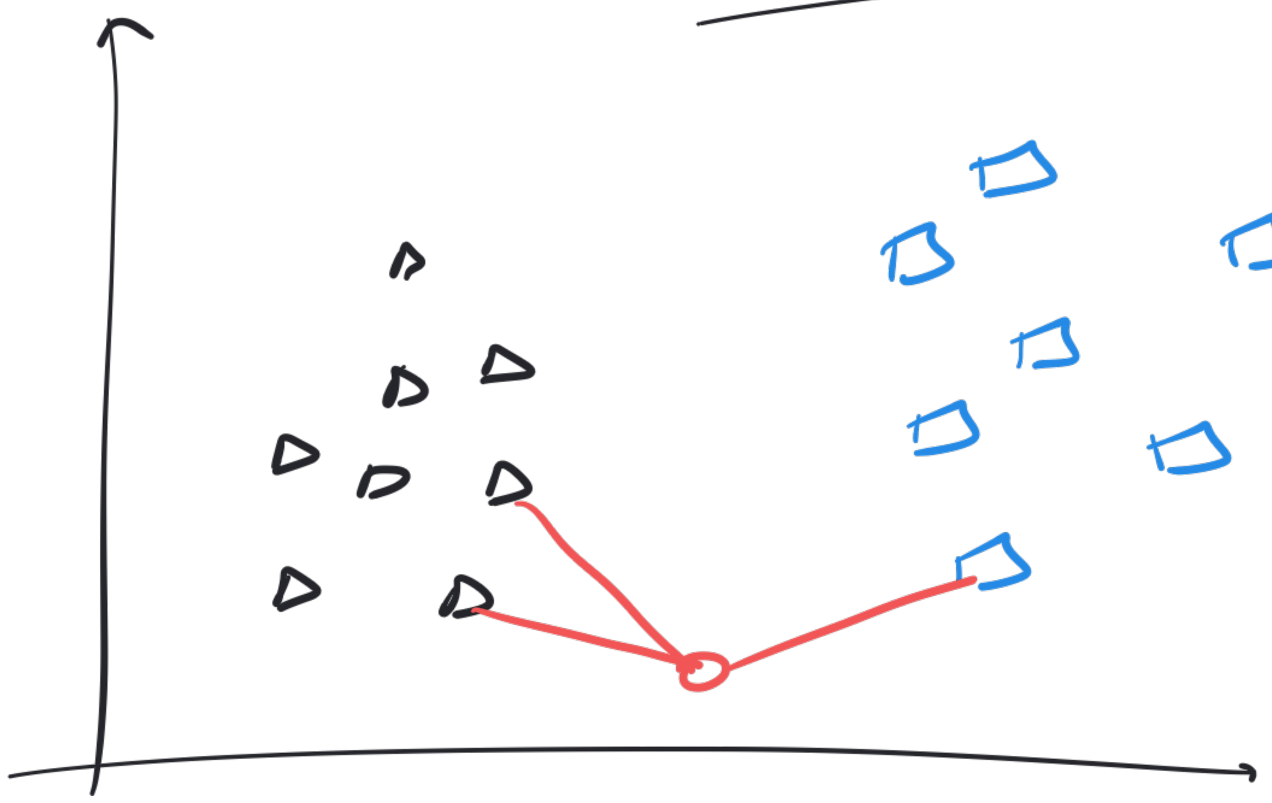
K-NEAREST NEIGHBORS (K-NN)

→ NON-parametric classifiER.

→ cannot use standard regularization penalties such as L1 and L2 norm but there are other strategies.

→ Instance-based learning

TRAINING DATA



Δ – class 1

□ – class 2

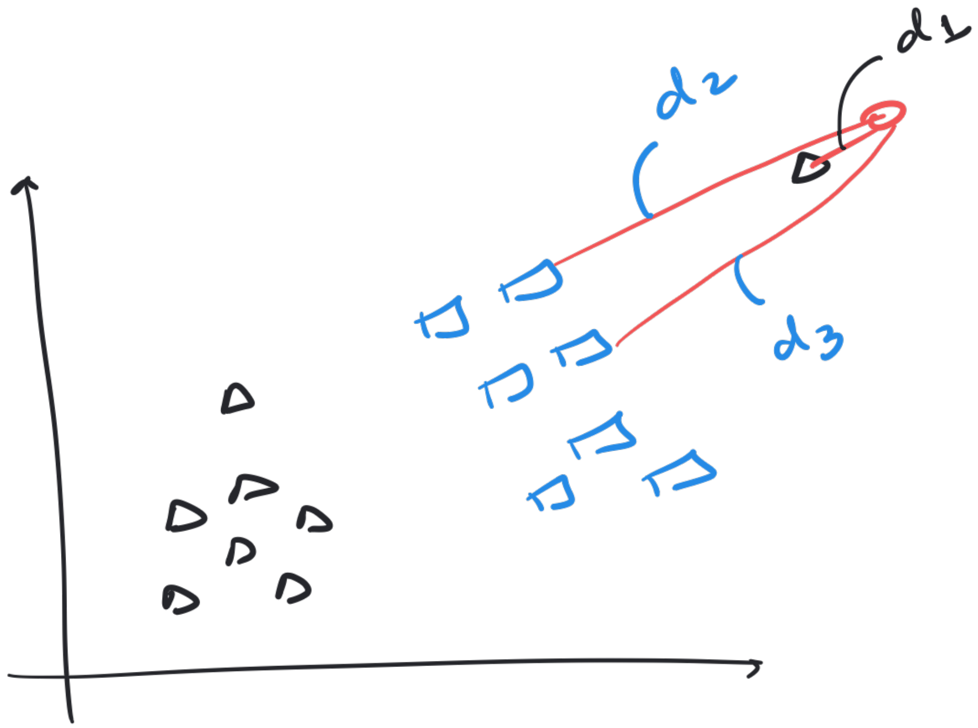
K = 3 (neighbors)

$\circ \in \Delta$

K-NN uses <u>majority vote</u> to assign class label.

TIE BREAKER SCHEMES:

1) pick one at random (flip a coin)
2) Assign label corresponding to the closest neighbor.
3) pick the class with largest prior probability
4) pick the class w/ <u>most</u> associated <u>risk</u>.



△ – class 1

▭ – class 2

○ – test point

K = 3

$$\boxed{\text{Weighted KNN}}$$

↳ sensitive to outliers

↳ prone to overfitting

$$\text{score}(\triangle) = \frac{1}{d_1}$$

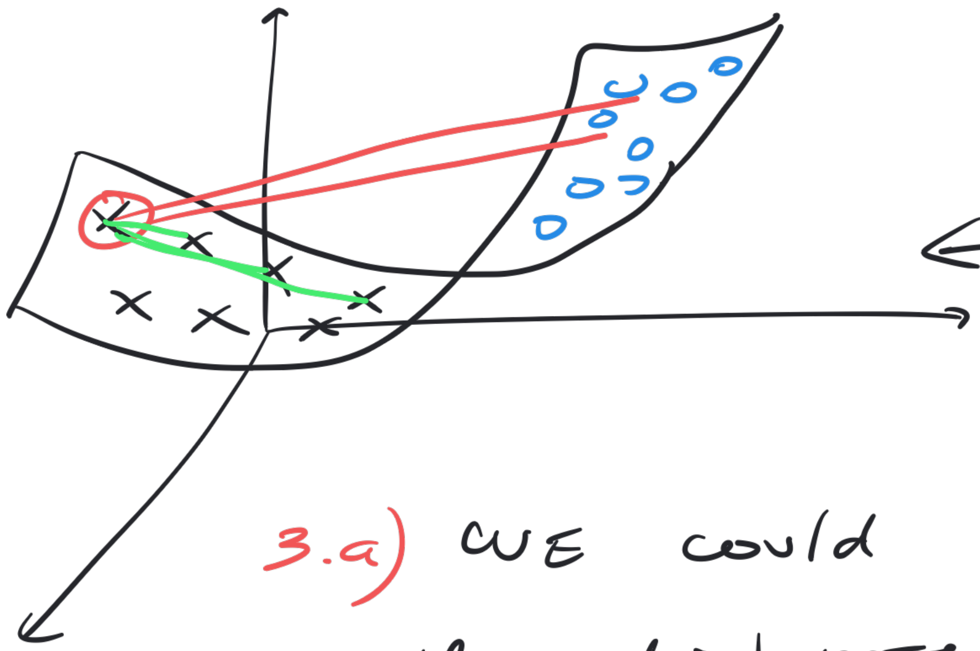
$$\text{score}(\square) = \frac{1}{d_2} + \frac{1}{d_3}$$

$\circ \in \triangle$ because $\text{score}(\triangle) > \text{score}(\square)$

OBSERVATIONS:

1) KNN is more sensitive to the Curse of Dimensionality

2) scaling is crucial!



3) For small K value in this manifold example, KNN provides a better generalization.

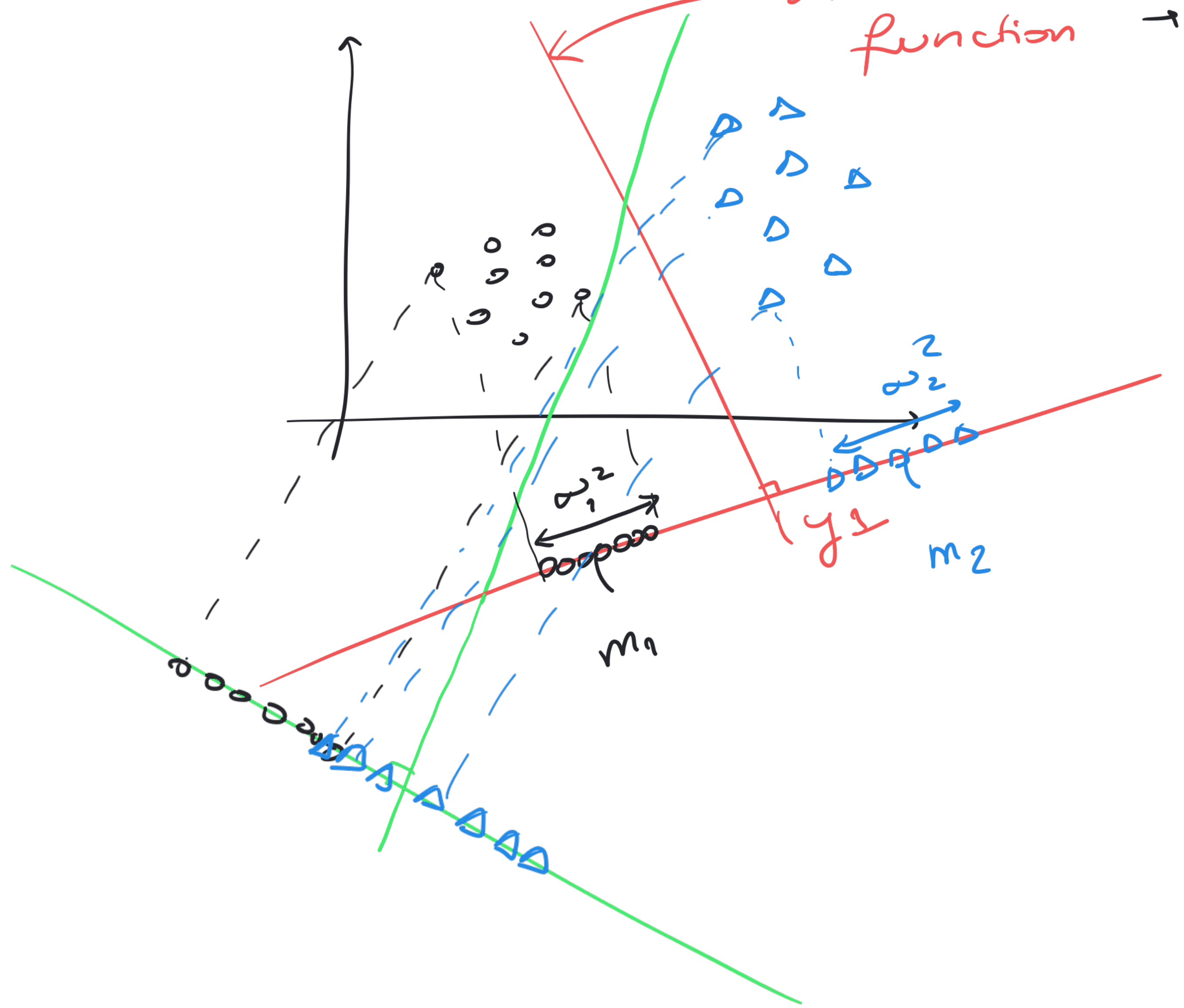
3.a) we could instead use shortest path distances.

4) K ≡ # neighbors is the hyperparameter. Must be tuned w/ cross-validation.

Fisher's Linear Discriminant Analysis

or LDA or FLDA

↳ linear discriminative classifier.



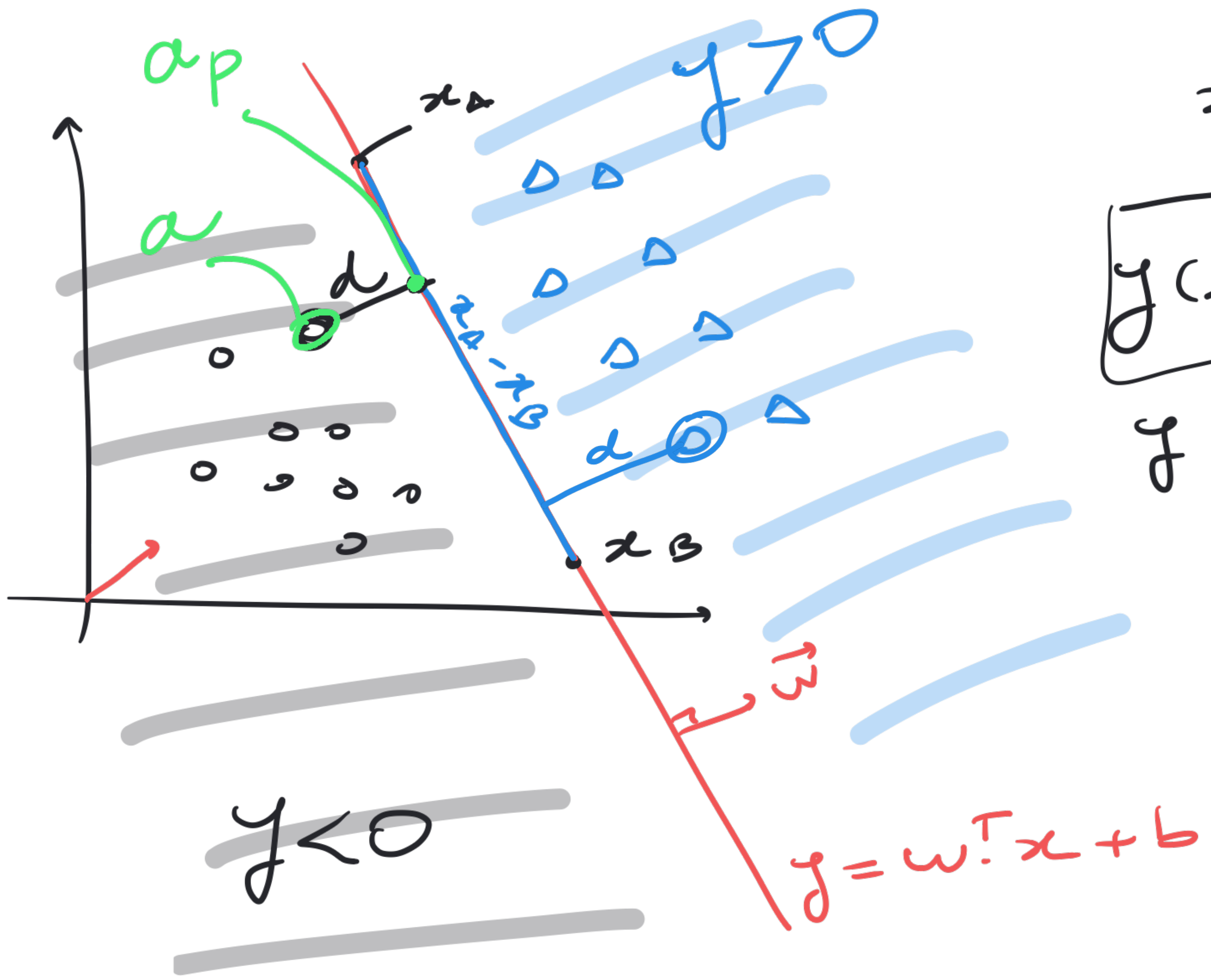
TRAINING → DATASET: $\{(x_i, t_i)\}_{i=1}^{N}$

$t_i \in \{1, 2, \dots, K\}$

Discriminant function:

$$\boxed{y = w^T x + b}$$

Goal: find w and b such that class projection variance is small and class projection mean-difference is large



$x_A \neq x_B \neq \vec{0}$, $\vec{w} \neq \vec{0}$

$$\boxed{y(x_A) = w^T x_A + b = 0}$$

$$y(x_B) = w^T x_B + b = 0$$

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \quad x_A = \begin{bmatrix} x_{A,1} \\ x_{A,2} \end{bmatrix}$$

$$b \in \mathbb{R}$$

$$w^T x_A + b = w^T x_B + b \iff w^T (x_A - x_B) = 0$$

$$\Rightarrow w \perp (x_A - x_B)$$

"orthogonal" or perpendicular

$a \equiv$ data sample

$a_p \equiv$ projection of a onto the discriminant fct y.

$d \equiv$ displacement ("distance") to the y.

$$a = a_p + d \cdot \underbrace{\frac{w}{\|w\|}}_{\text{direction of vector } \vec{w}}$$

Reminder:

$\|w\| = (w^T w)^{1/2}$

Left-multiply by w^T:

$$w^T \cdot a = w^T \cdot a_p + d \cdot \frac{w^T \cdot w}{\|w\|}$$

$$w^T \cdot a = w^T \cdot a_p + d \cdot \|w\|$$

Adding intercept b on both sides:

$$\underbrace{w^T \cdot a + b}_{y(a)} = \underbrace{w^T \cdot a_p + b}_{y(a_p) = 0} + d \cdot \|w\|$$

$$\Leftrightarrow \boxed{d = \frac{y(a)}{\|w\|}}$$

$$y = w^T \phi(x) + b$$