What is a Good Distance Measure?

Application dependent…but, metric properties help indexing, search, and retrieval.

A metric distance, Δ, must satisfy the following conditions:

- self-minimality: $\Delta(s,s) = 0$
- minimality $\Delta(s_1,s_2) >= \Delta(s_1,s_1)$
- symmetry $\Delta(s_1, s_2) = \Delta(s_2, s_1)$
- triangular inequality $\Delta(s_1,s_2) + \Delta(s_2,s_3) >= \Delta(s_1,s_3)$

Self-Minimality and Minimality

Self-minimality:

- $\Delta(s,s) = 0$
- Ensures that a given object matches itself perfectly

minimality

- $\Delta(s_1,s_2) >= \Delta(s_1,s_1)$
- Ensures that a no other object can match the given object better than itself

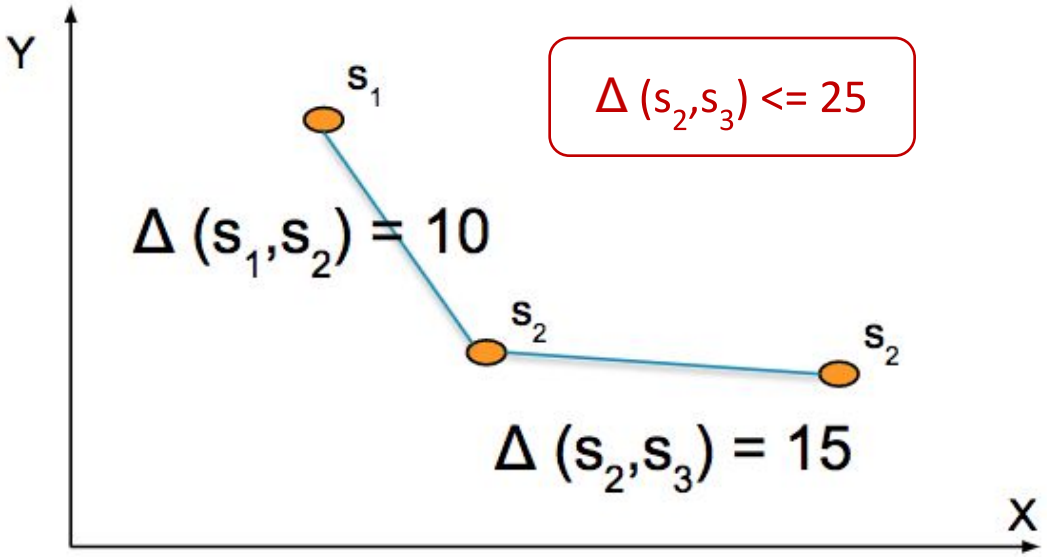
Symmetry

Symmetry:

- $\Delta(s_1, s_2) = \Delta(s_2, s_1)$
- Ensures that if a given object s_1 is matching another object s_2, then s_2 is equally matching s_1

Triangular Inequality:

- $\Delta(s_1,s_2) + \Delta(s_2,s_3) >= \Delta(s_1,s_3)$
- Enables effective pruning of the search space during retrieval



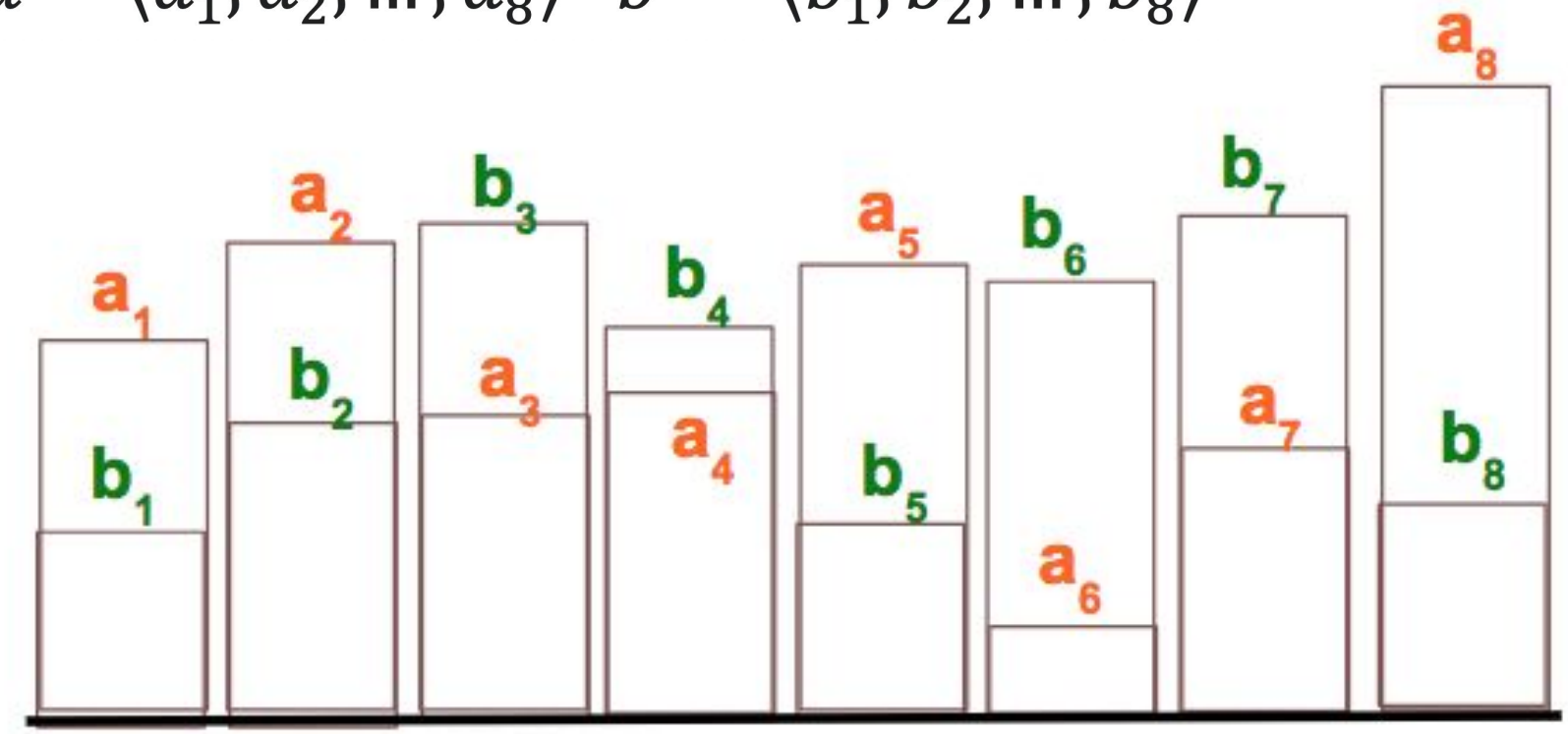
P-norms are Metric

- 1-norm, L1-metric $\left(\sum_{i=1..d} |v_{1,i} - v_{2,i}| \right)$
- 2-norm, L2-metric $\left(\sum_{i=1..d} |v_{1,i} - v_{2,i}|^2 \right)^{1/2}$
- ∞-norm, L∞-metric $\max_{i=1..d} |v_{1,i} - v_{2,i}|$

Are there other Similarity Measures?

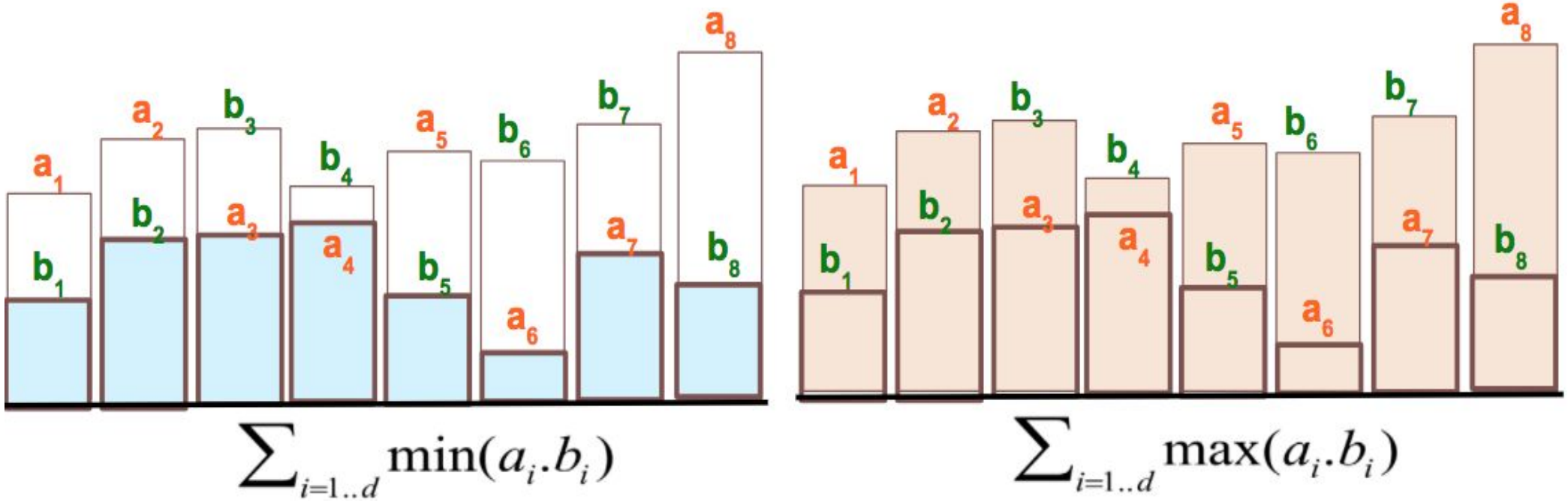
Consider two vectors

$$\vec{a} = \langle a_1, a_2, \ldots, a_8 \rangle \quad \vec{b} = \langle b_1, b_2, \ldots, b_8 \rangle$$



Intersection Similarity

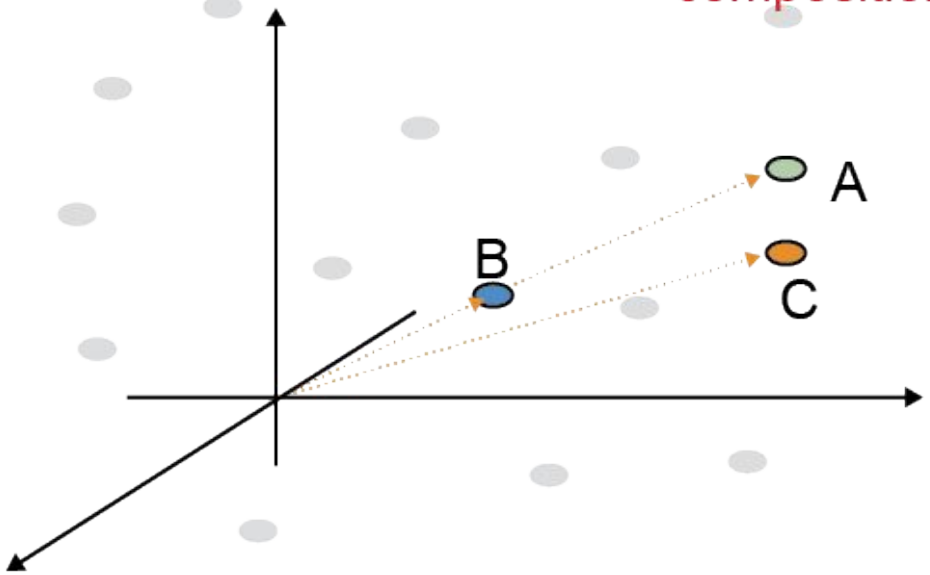
$$\mathrm{sim}_{\mathrm{int}}\left(\vec{a},\vec{b}\right)=\frac{\sum_{i=1..d}\min(a_i.b_i)}{\sum_{i=1..d}\max(a_i.b_i)}$$



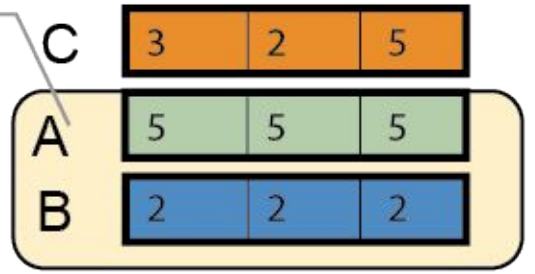
Angle based similarity measures

If we use the angle as a similarity measure, then A is more similar to E than F

$\cos(\widehat{AB}) > \cos(\widehat{AC})$



Similar composition



Angle-based Measures

- Given $\vec{a} = \langle a_1, a_2, \ldots, a_n \rangle \quad \vec{b} = \langle b_1, b_2, \ldots, b_n \rangle$

- Cosine similarity

$$sim_{cos}(\vec{a}, \vec{b}) = \cos(\vec{a}, \vec{b}) = \frac{\vec{a}.\vec{b}}{|\vec{a}||\vec{b}|}$$

- Dot product similarity

$$sim_{dot}(\vec{a}, \vec{b}) = \vec{a}.\vec{b} = \sum_{i=1}^{n} \boldsymbol{a_i b_i}$$

Cosine and dot product are the same if the vectors are unit length

Other Commonly Used Similarity / Distance Measures

- Pearson's correlation (similarity measure)
  - Linear correlation (the strength of linear association) among the corresponding components of two vectors
- KL-Divergence (distance measure)
  - How one vector (interpreted as a probability distribution) diverges from the other
- Earth-movers distance (distance measure)
  - How one vector (interpreted as a probability distribution) diverges from the other