The background features a complex geometric pattern of thin, intersecting lines in shades of brown and grey, creating a mesh-like effect. Overlaid on this is a semi-transparent white banner containing the title. To the left of the banner, there is a small inset image showing a scatter plot with orange and brown dots, and a grid of pink and white squares.

# **Proximity Measure between Two Vectors: Cosine Similarity**



# Cosine Similarity of Two Vectors

- A **document** can be represented by a bag of terms or a long vector, with each attribute recording the *frequency* of a particular term (such as word, keyword, or phrase) in the document

Document	teamcoach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	2	0	0
Document2	3	0	2	0	1	1	1	0	1
Document3	0	7	0	2	1	0	3	0	0
Document4	0	1	0	0	1	2	0	3	0

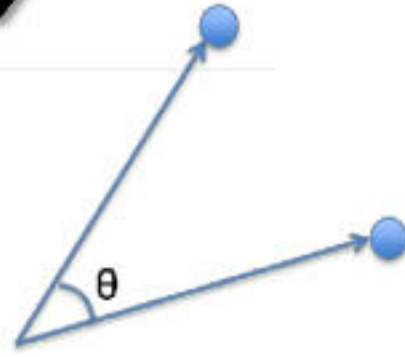
- Other vector objects: Gene features in micro-arrays
- Applications: Information retrieval, biologic taxonomy, gene feature mapping, etc.
- Cosine measure: If  $d_1$  and  $d_2$  are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\| \times \|d_2\|}$$

where  $\bullet$  indicates vector dot product,  $\|d\|$ : the length of vector  $d$

# Example: Calculating Cosine Similarity

□ Calculating Cosine Similarity:  $\cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\| \times \|d_2\|}$   $\text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$



where  $\bullet$  indicates vector dot product,  $\|d\|$ : the length of vector  $d$

□ Ex: Find the **similarity** between documents 1 and 2.

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) \quad d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

□ First, calculate vector dot product

$$d_1 \bullet d_2 = 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 1 + 2 \times 1 + 0 \times 0 + 0 \times 1 = 25$$

□ Then, calculate  $\|d_1\|$  and  $\|d_2\|$   $= \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$

$$\|d_1\| = \sqrt{5 \times 5 + 0 \times 0 + 3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0} = 6.481$$

$$\|d_2\| = \sqrt{3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 1 \times 1 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 1} = 4.12$$

□ Calculate cosine similarity:  $\cos(d_1, d_2) = 25 / (6.481 \times 4.12) = 0.94$