Lecture 16: What are Embeddings

What Exactly Is an Embedding?

- Computers think in numbers.
- **Embeddings** turn text/images into lists of numbers (**vectors**) so AI can understand meaning, not just exact words.
- A vector of floating-point numbers (not a single number).
- Words/sentences with similar meaning → vectors that are close to each other.
- Example: dog ↔ cat (close), but dog ↔ laptop (far).

Where Are Embeddings Used?

- **Search** (find related results even if wording differs)
- **Clustering & Classification** (group similar things, label them)
- **Recommendations** (movies/books/products)
- **Anomaly Detection** (spot "odd" behavior)
- **Diversity Measurement** (avoid near-duplicates)
- Everyday cases: **spam filtering**, **semantic search**, etc.

How Does "Closeness" Work?

- Each item (e.g., "dog," "cat," "Java", "laptop," "Earth," "Saturn") gets a vector
- If two vectors are near in space, AI treats them as **related**.
- A quick 2D/3D plot helps intuition, but real models use **high dimensions**.

Dimensionality

- More dimensions → more nuance.
- Common sizes (from lecture examples)
    - **text-embedding-3-small** → **1536** dimension
    - **text-embedding-3-large** → **3072** dimensions
- Some models let you **limit dimensions** (trade accuracy for speed/storage).

Key Takeaways:

- **Definition: A** Numeric vector that captures meaning.
- **Use cases:** Search, clustering, recommendations, and anomaly detection.
- **Key property:** Distance ≈ semantic similarity.
- **Reality:** High-dimensional (e.g., 1536/3072)