Vector Databases & Embeddings Explained

1. What is an Embedding?

An embedding is a numerical representation of data (text, image, audio, or object) in the form of a vector (list of numbers). These vectors capture the semantic meaning of the data so that similar items have similar vectors.

2. Why Embeddings are Needed

Computers cannot understand raw text or images directly. Embeddings convert complex data into numbers that machine learning models can process. This enables similarity search, clustering, retrieval, and recommendation systems.

3. Types of Embeddings

- **Word Embeddings:** Convert words or sentences into vectors (e.g., Word2Vec, OpenAI text embeddings).
- **Image Embeddings:** Convert images into vectors using vision models like CLIP or ResNet.
- **Multimodal Embeddings:** Handle text + image together in the same vector space.

4. Example: Text to Vector

Sentence: "I love artificial intelligence"
Vector: [0.021, -0.44, 0.91, ...]
Similar sentences will have vectors close to each other in vector space.

5. What is a Vector Database?

A vector database is a specialized database designed to store, index, and search high-dimensional vectors efficiently. Instead of traditional SQL queries, it performs similarity search using distance metrics.

6. Popular Vector Databases

- Pinecone
- FAISS
- ChromaDB
- Weaviate
- Milvus

7. How Vector Search Works

Distance metrics used:

- Cosine Similarity
- Euclidean Distance
- Dot Product

8. Data → Vector → Database Flow

Step 1: Raw data (text, image, pdf)
Step 2: Embedding model converts data to vector
Step 3: Vector stored in vector database
Step 4: Query converted to vector
Step 5: Database finds nearest vectors (semantic search)

9. Example Use Case: RAG (Retrieval-Augmented Generation)

- Documents are embedded and stored in vector DB

- User query is embedded
- Relevant documents retrieved via similarity search
- LLM generates answer using retrieved context

10. Real World Applications

- Chatbots with memory
- Semantic search engines
- Recommendation systems
- Document Q&A; systems
- Image similarity search

11. Key Takeaway

Embeddings bridge the gap between human data and machine understanding. Vector databases make it possible to store and search this knowledge efficiently at scale.