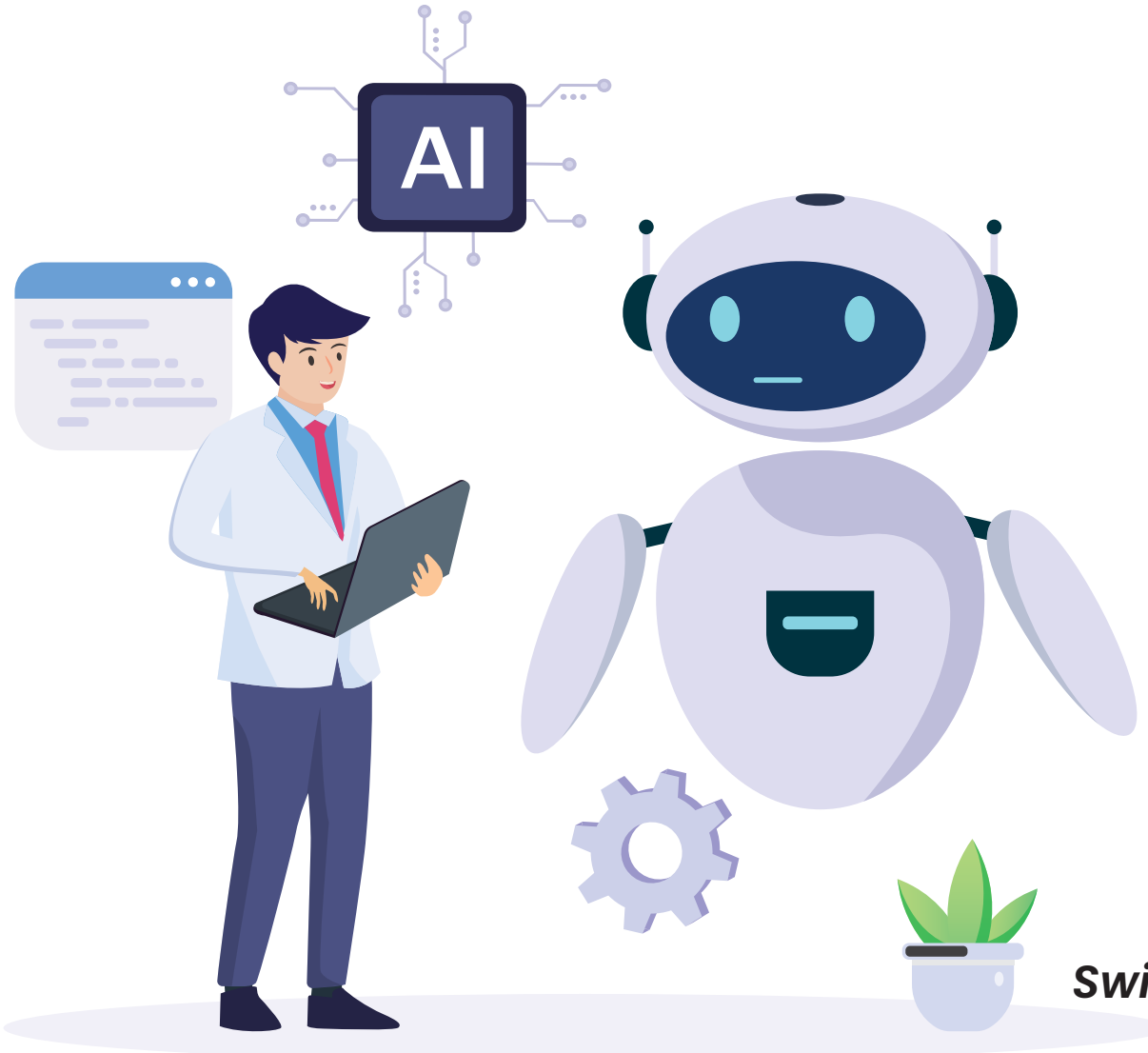
STATFUSIONAI

ESSENTIAL VECTOR DATABASE TERMINOLOGY FOR AI ENGINEERS



Swipe ▶

VECTOR DATABASE DICTIONARY

Vector Database

Database designed to store, manage, and query vector embeddings, enabling similarity searches across high-dimensional data.

Embedding

Numerical vector representation of data in a high-dimensional space where semantic similarity corresponds to vector proximity.

Similarity Search

Finding vectors closest to a query vector based on distance metrics like cosine similarity or Euclidean distance.

Vector Indexing:

Organizational structures that optimize vector search performance by reducing the need for exhaustive comparisons.

ANN (Approximate Nearest Neighbor)

Algorithms that find approximate nearest neighbors with high probability, trading perfect accuracy for speed.

Flat Index

Simple brute-force approach that compares query vector against every vector in the database; accurate but slow for large datasets.

HNSW (Hierarchical Navigable Small World)

Graph-based indexing algorithm creating multiple layers of connections between vectors for efficient navigation during search.

IVF (Inverted File Index)

Partitions vectors into clusters, allowing queries to search only relevant clusters rather than the entire dataset.

PQ (Product Quantization)

Compression technique that divides vectors into subvectors and quantizes each separately, reducing memory requirements.

IVFPQ

Combines IVF clustering with PQ compression for efficient large-scale vector search.

Indexing

Process of adding vectors to the database and organizing them within the chosen index structure.

Querying

Searching for vectors similar to a given query vector, typically returning k-nearest neighbors.

Filtering

Constraining vector search results based on metadata attributes.

Reindexing

Rebuilding an index after significant changes to optimize performance.

Sharding

Distributing vector data across multiple machines to handle scale.

Recall

Percentage of true nearest neighbors found by an approximate search algorithm.

Latency

Time required to complete a single search query.

Throughput

Number of queries processed per second.

QPS (Queries Per Second)

Measure of search performance under concurrent load.

Index Build Time

Duration required to construct the vector index.

POPULAR VECTOR DATABASES

FAISS

Facebook AI Similarity Search library optimized for efficient similarity search.

PINECONE

Managed vector database service with real-time updates.

MILVUS

Open-source vector database built for scalability.

WEAVIATE

Vector search engine with GraphQL interface.

QDRANT

Vector database focusing on extended filtering and payload storage.

CHROMA

Open-source embedding database designed for RAG applications.

MOST IMPORTANT CONCEPTS FOR INTERVIEWS

1. **Vector Embeddings:** Understanding how data is transformed into vectors that capture semantic meaning
   - Definition: Numerical representations of data in high-dimensional space where similar items are positioned closer together

2. **HNSW Indexing:** The most widely used high-performance vector index structure
   - Definition: Hierarchical graph-based index that creates multiple layers of connections between vectors for efficient navigation during search

3. **ANN vs. Exact Search:** Trade-offs between speed and accuracy
   - Definition: Approximate methods sacrifice perfect accuracy for dramatically improved search speed on large datasets

4. **Similarity Metrics:** Knowing when to use cosine similarity vs. Euclidean distance
   - Definition: Mathematical measures that quantify the similarity between vectors, with different applications depending on data characteristics

5. **Filtering:** How to combine vector search with metadata filtering
   - Definition: Technique to narrow vector search results based on additional metadata attributes beyond vector similarity

6. **Recall vs. Latency:** Understanding the performance trade-offs
   - Definition: Fundamental trade-off where increasing search accuracy (recall) typically comes at the cost of increased search time (latency)

7. **Quantization:** Compression techniques for efficient storage and retrieval
   - Definition: Process of mapping continuous vector values to discrete values to reduce memory requirements and improve search speed

8. **Sharding and Scaling:** How vector databases handle large-scale deployments
   - Definition: Techniques for distributing vector data across multiple machines to maintain performance at scale

9. **Flat Index:** Understanding the baseline for comparison
    - Definition: Exhaustive search method that compares query against every vector, providing perfect recall but scaling poorly with dataset size
10. **Reranking:** Two-stage retrieval for improved results
    - Definition: Process of retrieving a large initial set of candidates with ANN then refining results with a more accurate but computationally expensive model