Vector Databases

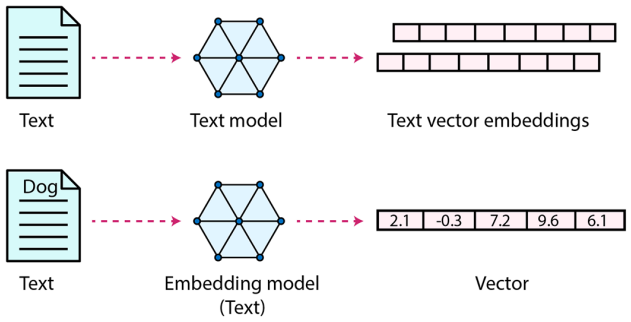
Vector databases, a newer NoSQL database, are rapidly becoming popular with the exponential increase in the use of Large Language Models (LLMs), such as OpenAI's GPT. But what is a vector database? Here's how IBM defines a vector database:

> A vector database is designed to store, manage, and index massive quantities of high-dimensional vector data efficiently.
> **Source:** https://www.ibm.com/topics/vector-database

Next, learn about data as vectors.

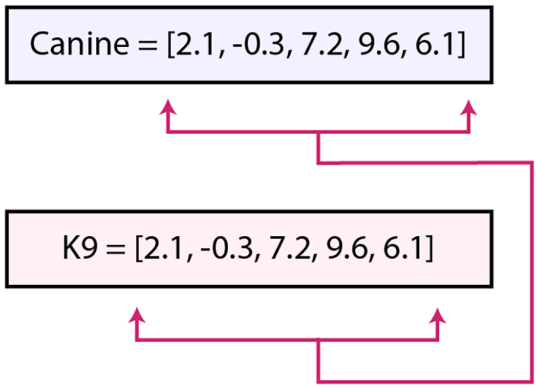
Vectors

You can transform text, images, audio, and video into vectors, also known as vector data, using embedding functions based on various methods, including machine learning models, word embeddings, and feature extraction algorithms.



For example, the vector representation of dog is [2.1,-0.3, 7.2, 9.6, 6.1]

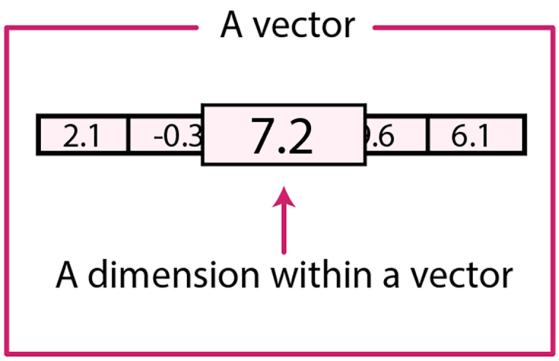
Similar words for dogs include the word *canine* or *K9*, so a vector database will identify both terms and include the same vector values.



Important! When words have relationships or similar contexts, but the meaning is not identical, these words have vectors that are closer together within the database that help identify the relationship.

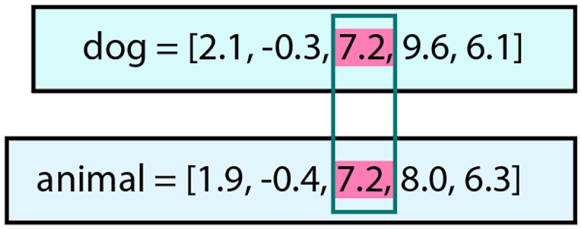
For example, you'll see the word animal represented as [1.9, -0.4, 7.2, 8.0, 6.3]

Each numeric value you see displayed inside of a vector is one dimension.



Vectors can have a few or thousands of dimensions, depending on the granularity and complexity of the classification required.

Returning to the example of a dog classifications, we know a dog is an animal, but not all animals are dogs. The relevance and relationships between the words "dog" and "animal" mean that these words will be much closer together as vector values and within the database itself. The following illustration shows the one dimension, dimension 7.2, that the word **dog** and **animal** share.



Next, explore why companies are moving to vector databases.

Vector Database benefits

In contrast to conventional techniques that involve querying databases for exact matches or predefined criteria, a vector database empowers you to discover the most similar or related data by considering their semantic or contextual significance.

In other words, unlike other database types that require an exact term search, you can use a vector database to conduct similarity searches and retrieve data according to their vector distance or likeness.

For example, you can use a vector database to perform the following tasks:

- Recommend TV shows to watch based on your current viewing habits.
- Locate related products based on the first product's features and ratings when shopping online.

Because related vector data exists mathematically closer to each other within the database, search and data delivery times are faster. So rather than having to perform additional analysis techniques to retrieve related data, the trained model and vector database delivers relevant search results faster.

Popular Vector Databases

Database offerings and their features are changing concurrently with the exponentially fast speed of AI development. Before selecting a vector database, you'll want to review its applicability to your data and LLM. Next, check out these currently popular vector databases:

Chroma

Chroma is an open source embedding database with which you can perform the following tasks:

- Store embeddings and their metadata
- Embed documents and queries
- Search embeddings

Pinecone

Pinecone provides long-term memory for high-performance AI applications. Pinecone emphasizes the following capabilities and features:

- Runs as a fully managed service
- Provides high scalability
- Provides real-time data ingestion
- Delivers low-latency search

Weaviate

Weaviate, an open-source vector database that stores data objects and vector embeddings from machine-learning models, is said to provide the following capabilities and features:

- Provides efficient similarity searches
- Scales to store and process billions of data objects
- Runs the GraphQL API
- Provides real-time updates

Recap

After completing this reading, you know that:

- Vector databases store, manage, and index massive quantities of high-dimensional vector data efficiently.
- You can transform text, images, audio, and video into vector data.
- Vector data consists of a series of numbers known as dimensions.
- Vector databases store Items with similar or "like" vector numbers closely together within the database
- Chroma, Pinecone, and Weaviate are three popular vector databases

Ready to learn more? Get in-depth information at the following IBM websites:

IBM.com: What is a vector database?
The IBM Engineering and Scientific Subroutine Library: Vectors

Congratulations! You have completed this lab and are ready for the next topic.

Author(s)

Patsy Kravitz