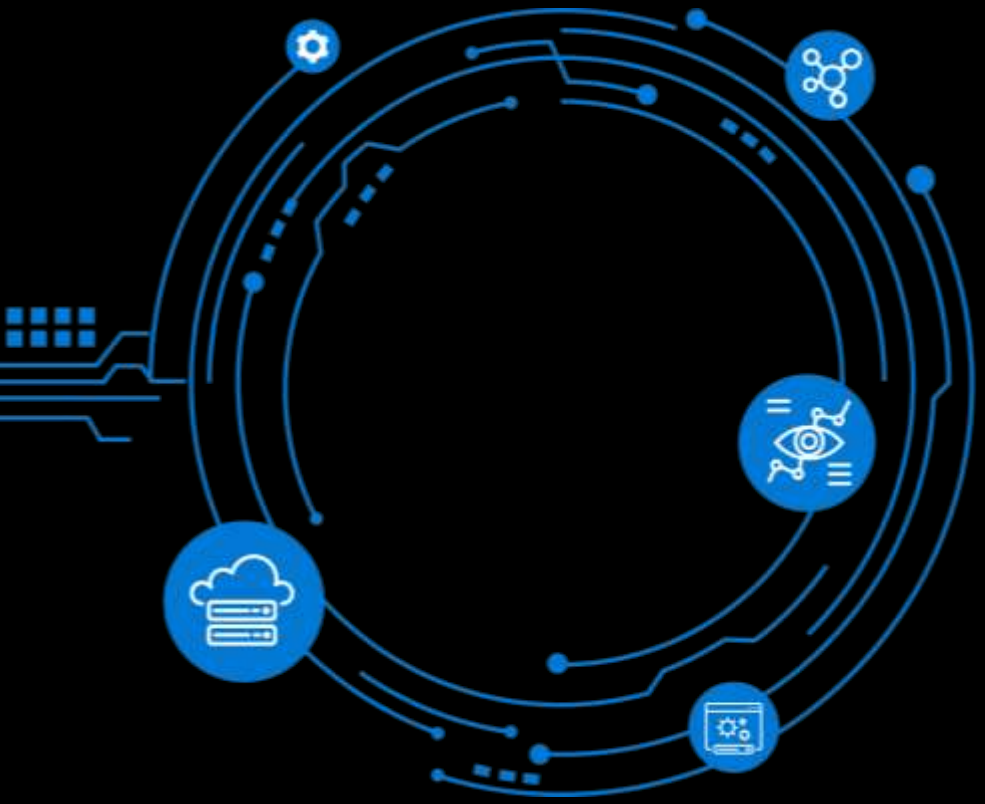
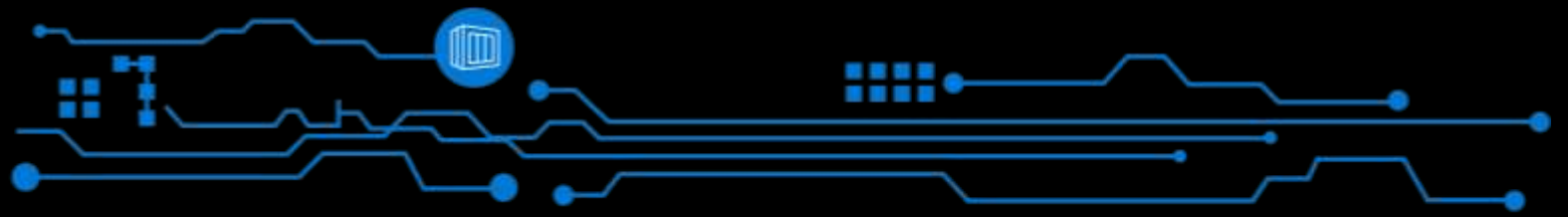
Microsoft





Vector search on embeddings for developers

Exploring **Azure Cosmos DB's** Vector search Capabilities

Divakar-Kumar

Agenda

- What
- How
- Why

About me

Cloud Architect @FlyersSoft | Microsoft MVP

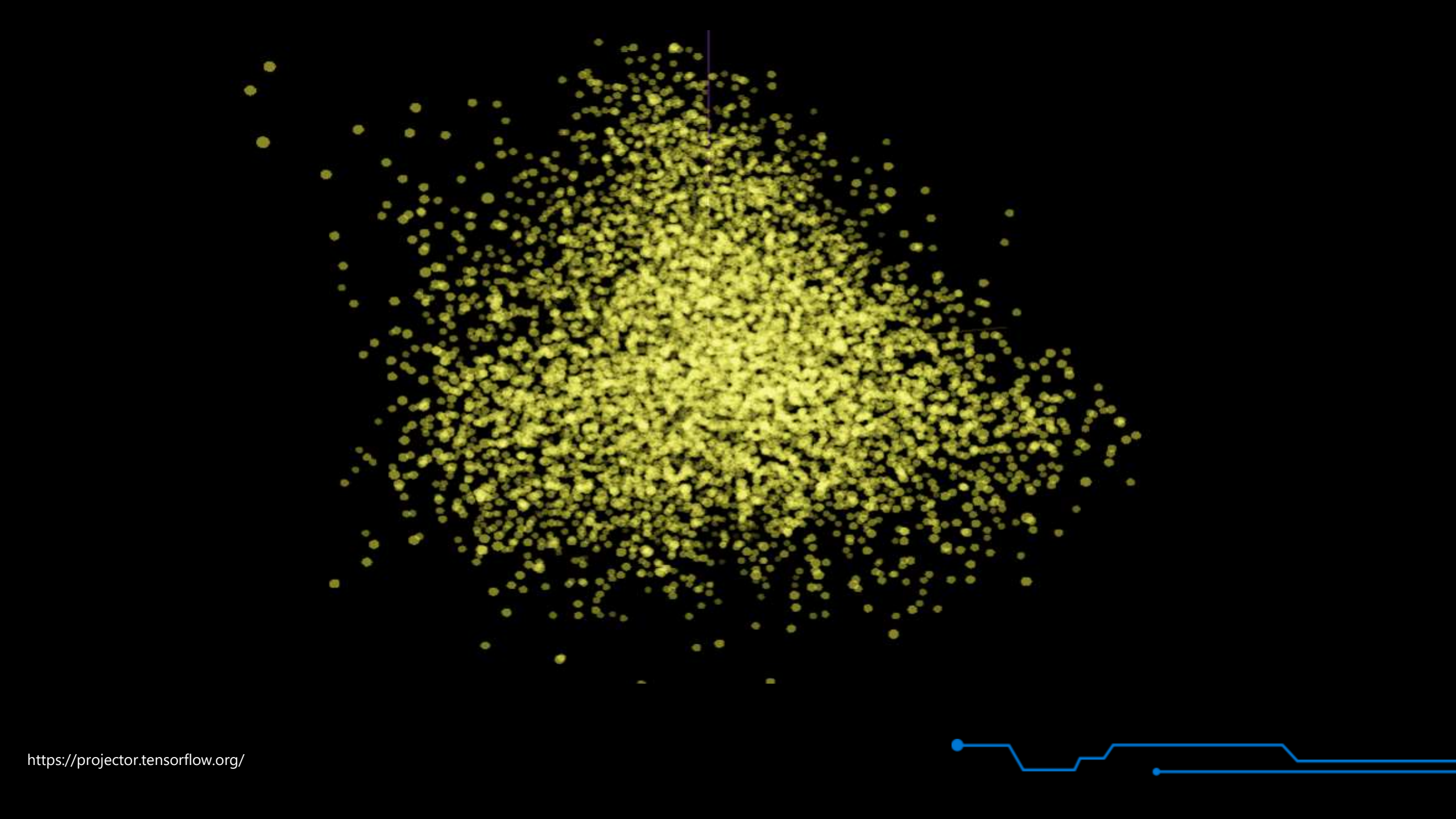
Divakar-Kumar

Divakar-Kumar

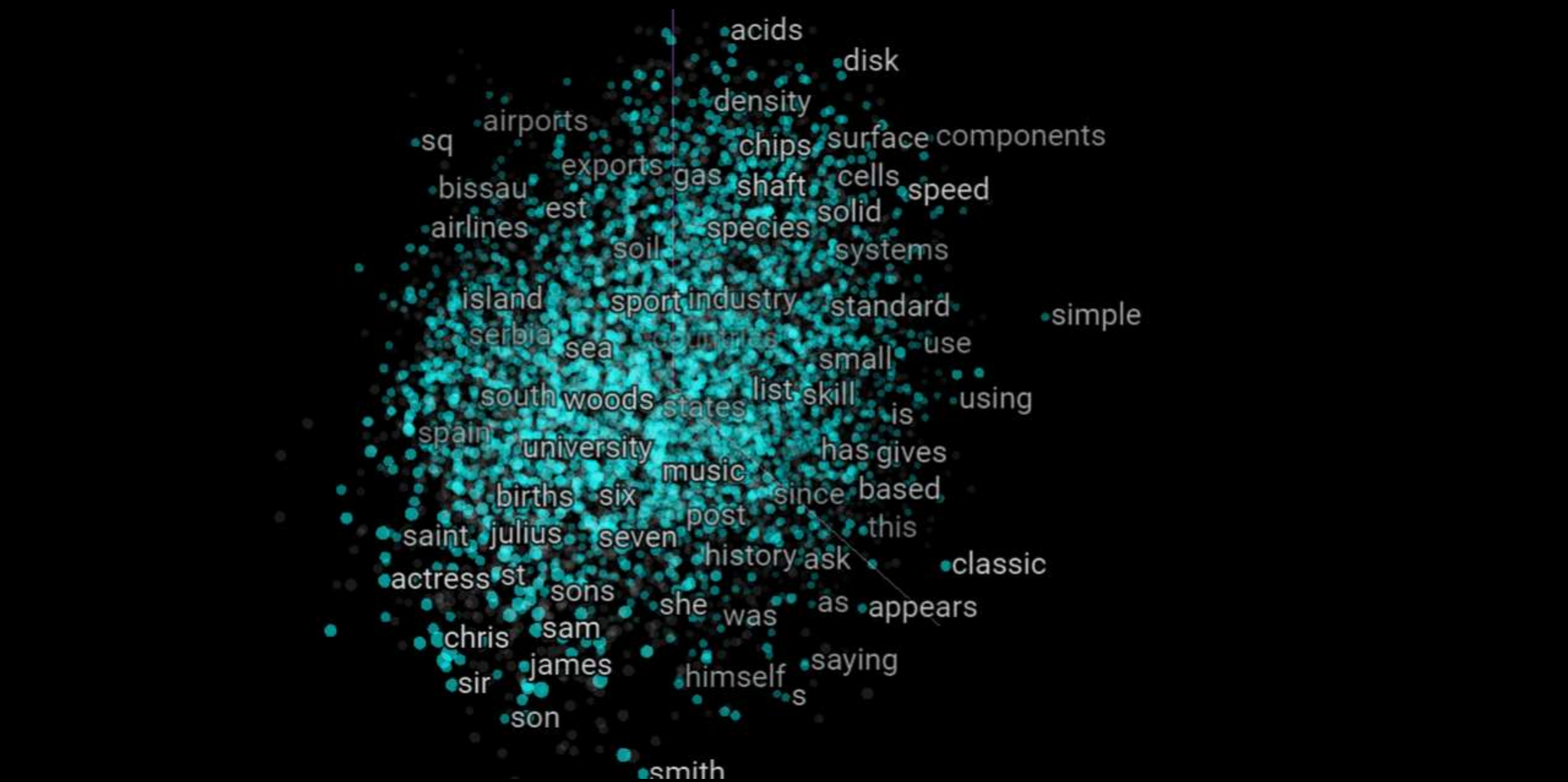
What is a vector database?

A vector database is a type of database that stores data as high-dimensional vectors, which contains hundreds of dimensions, and each dimension corresponds to a specific feature or property of the data object it represents.

Vector data usually generated by embedding function also known as vector embeddings. Vector database is different from vector search or vector index. It is a data management solution that enables scalability, perform backups , offer security features, storage and filtering.



https://projector.tensorflow.org/

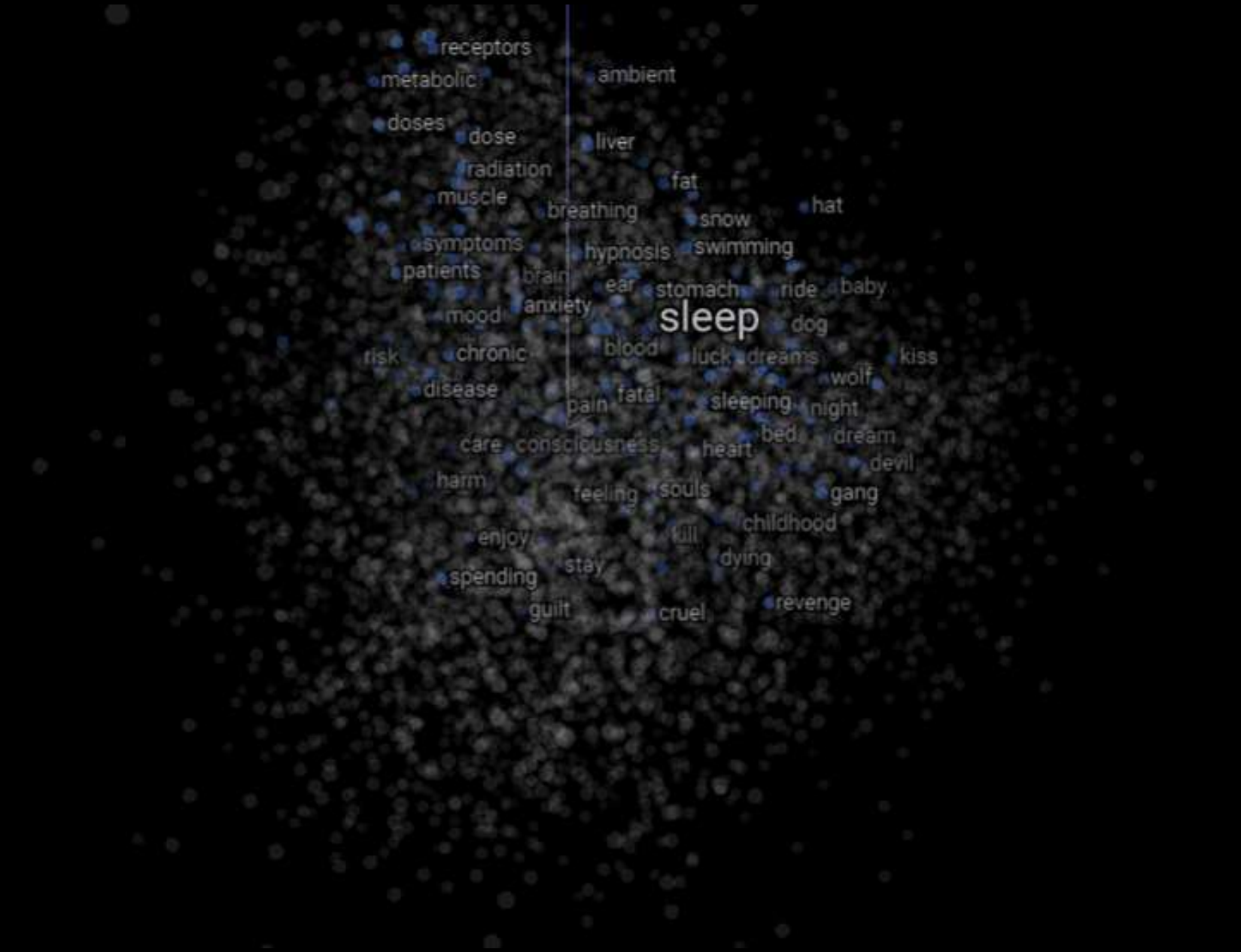


https://projector.tensorflow.org/

What are embeddings?

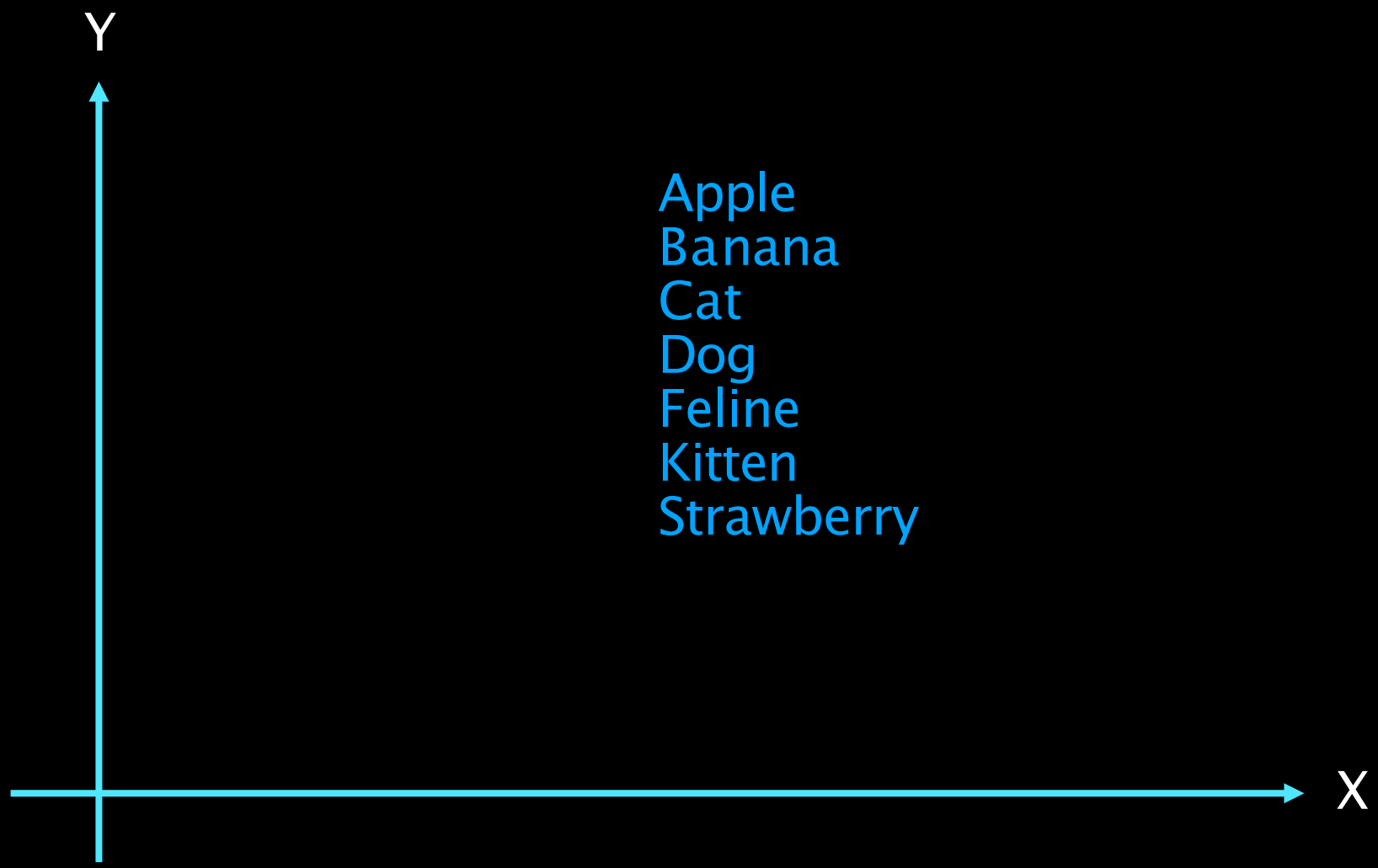
Embeddings are a technique for representing words or data as vectors in a high-dimensional space, allowing us to measure their relatedness.

In AI models, embeddings help computers understand the meaning and context of words or data. To use embeddings, the text is transformed into vectors, and a search query is converted into its vector representation to find similar vectors. This process resembles searching on a search engine, providing ranked matches based on similarity.



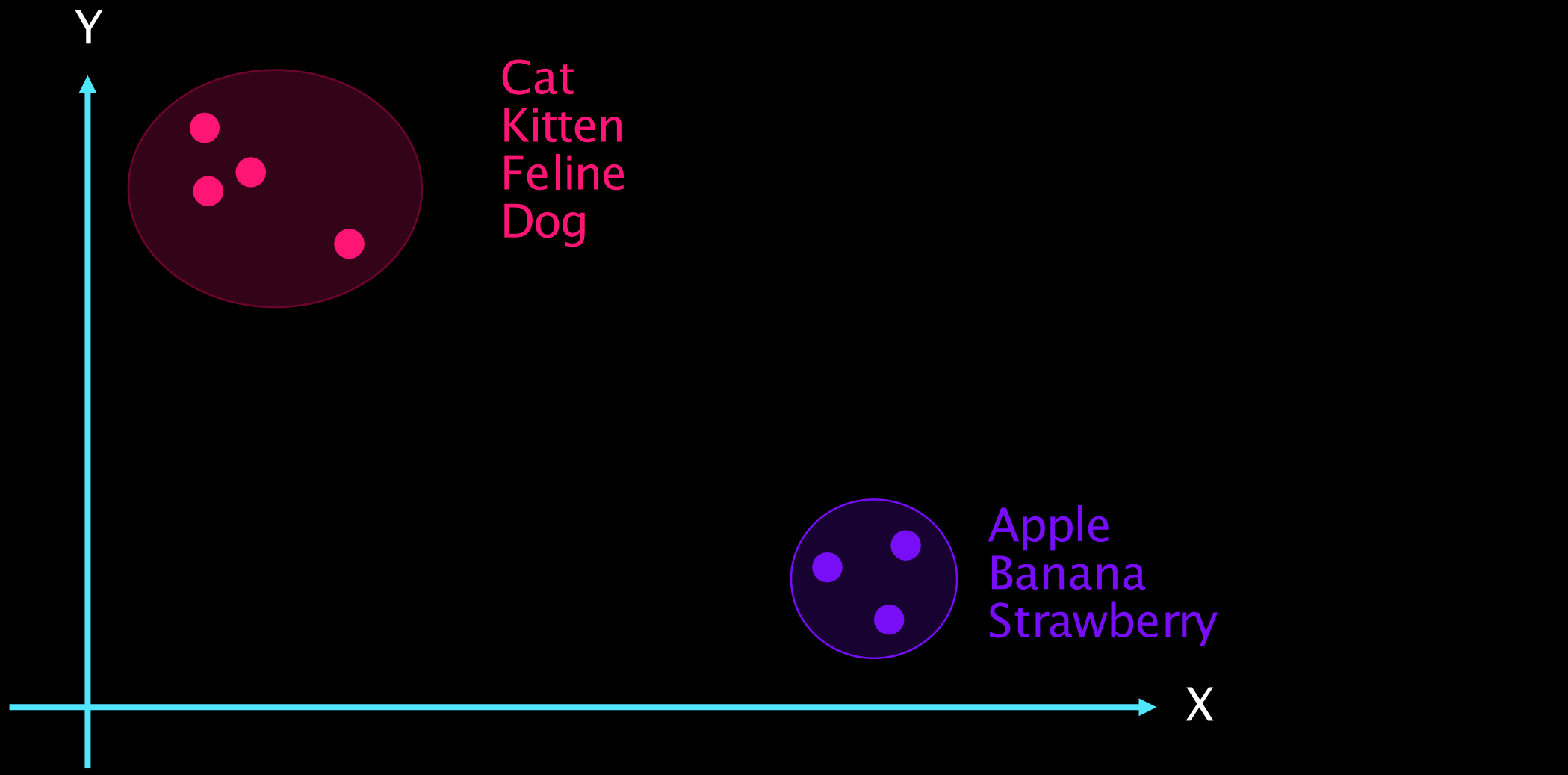
https://projector.tensorflow.org/

2-Dimensional space

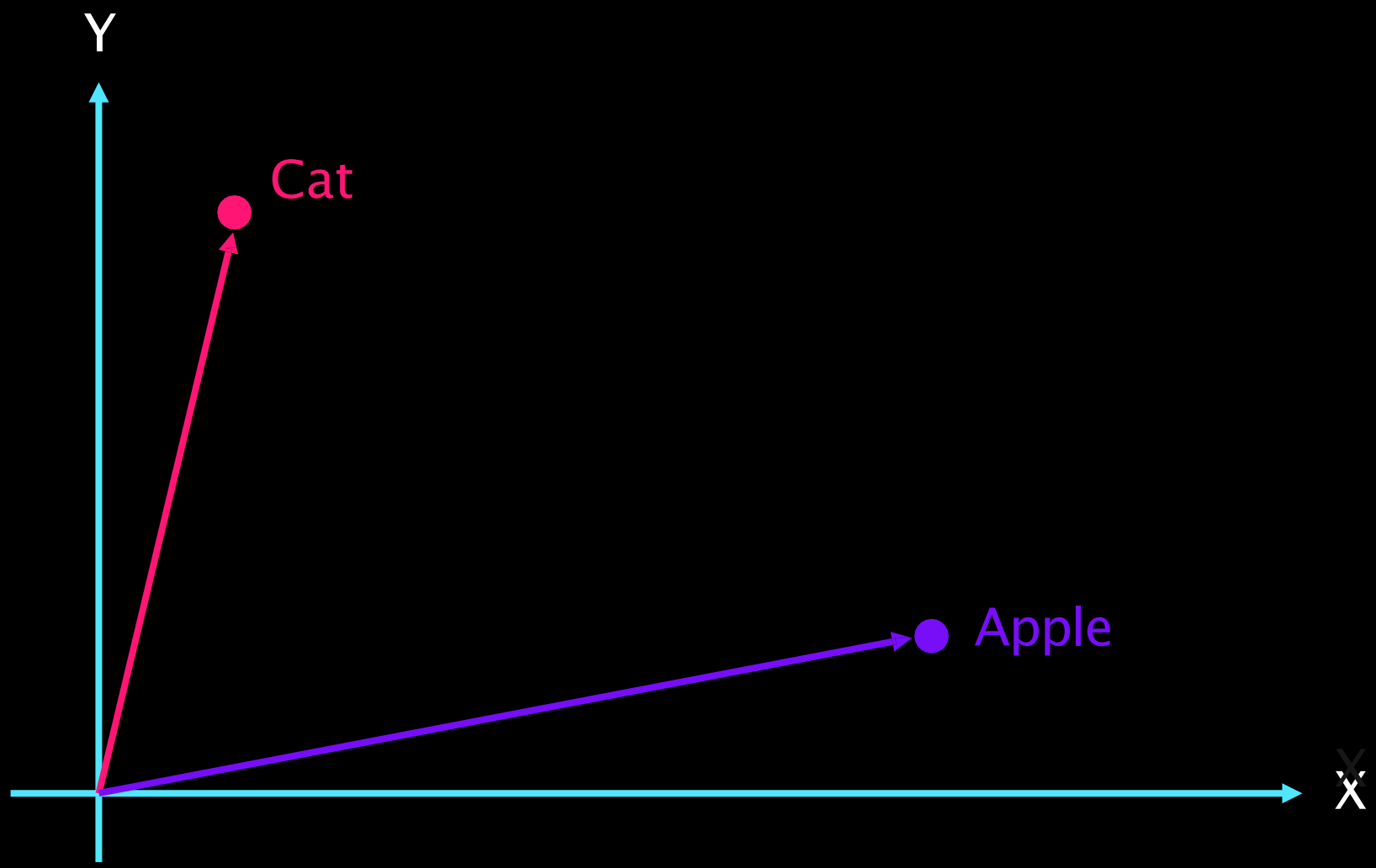


Apple
Banana
Cat
Dog
Feline
Kitten
Strawberry

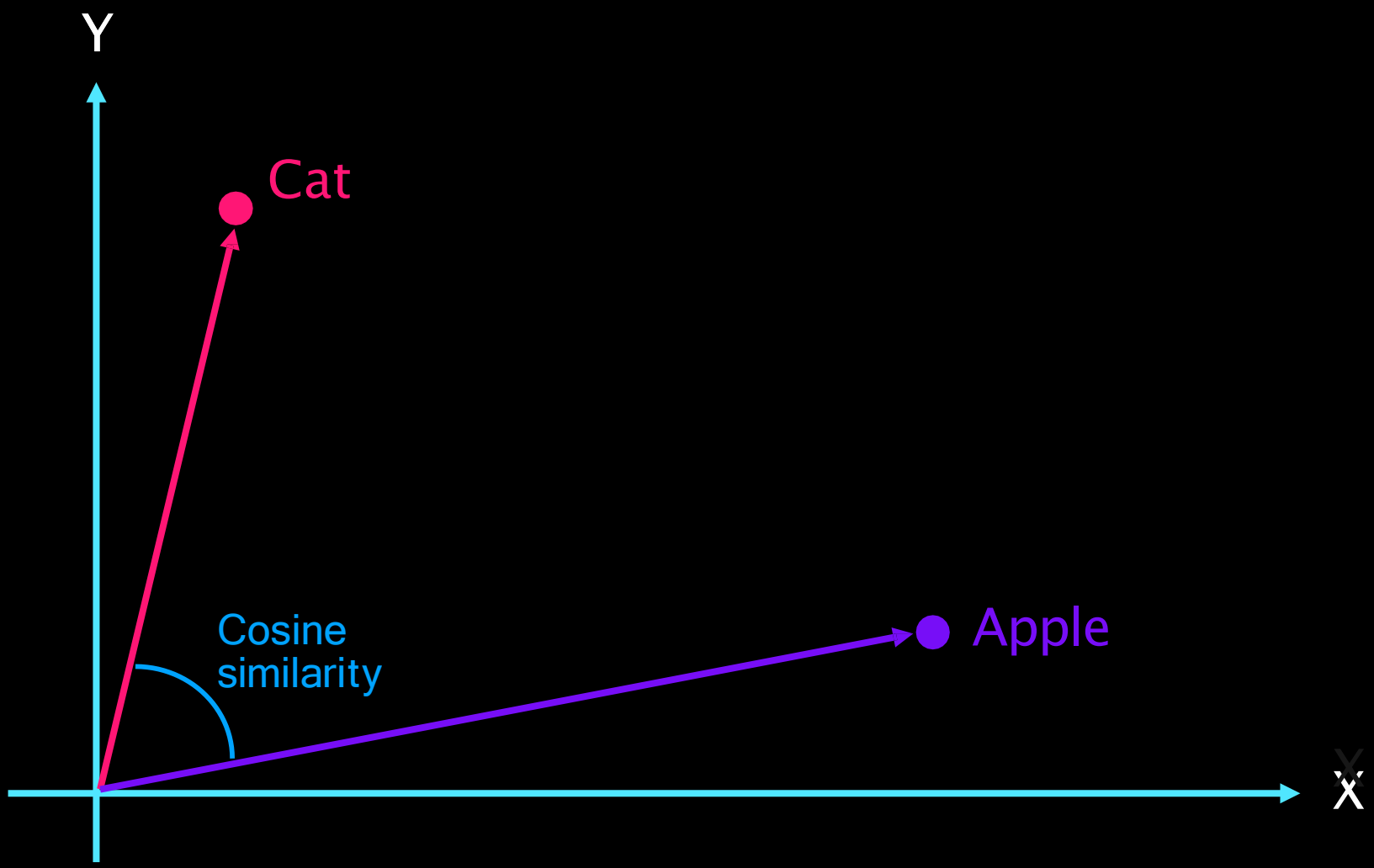
2-Dimensional space



2-Dimensional space



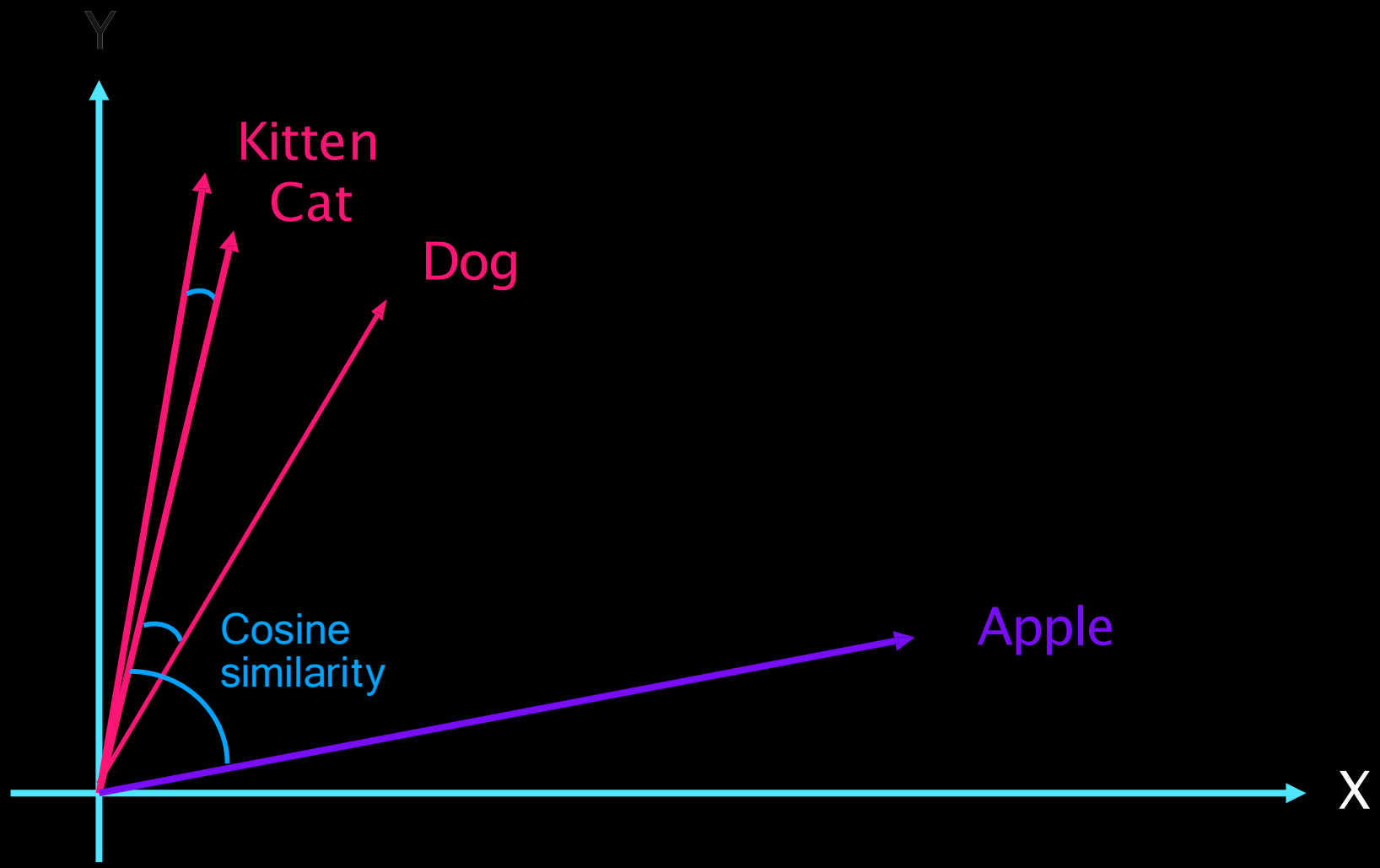
2-Dimensional space

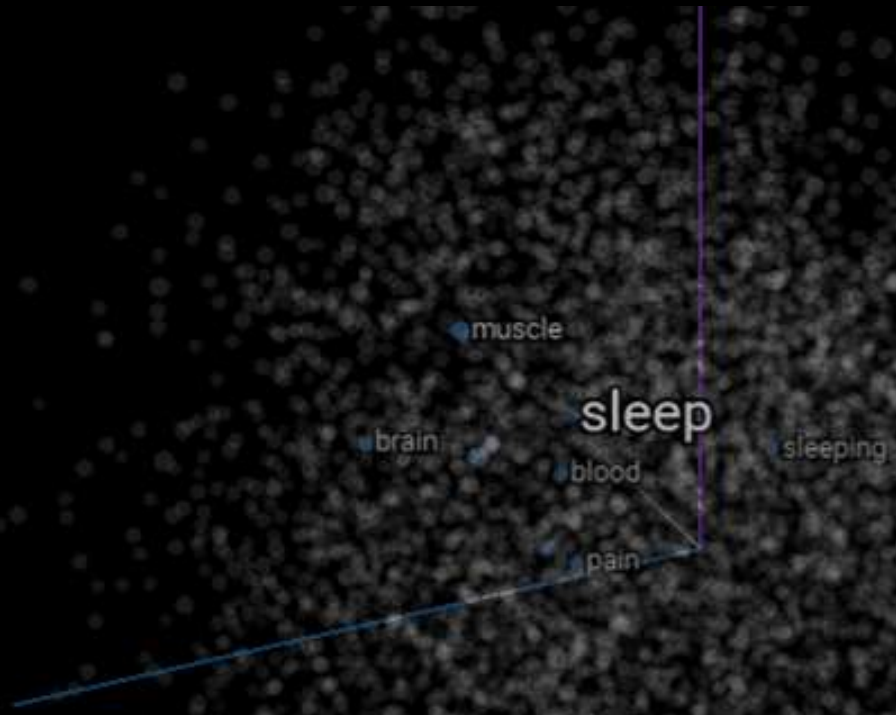


Distance metrics

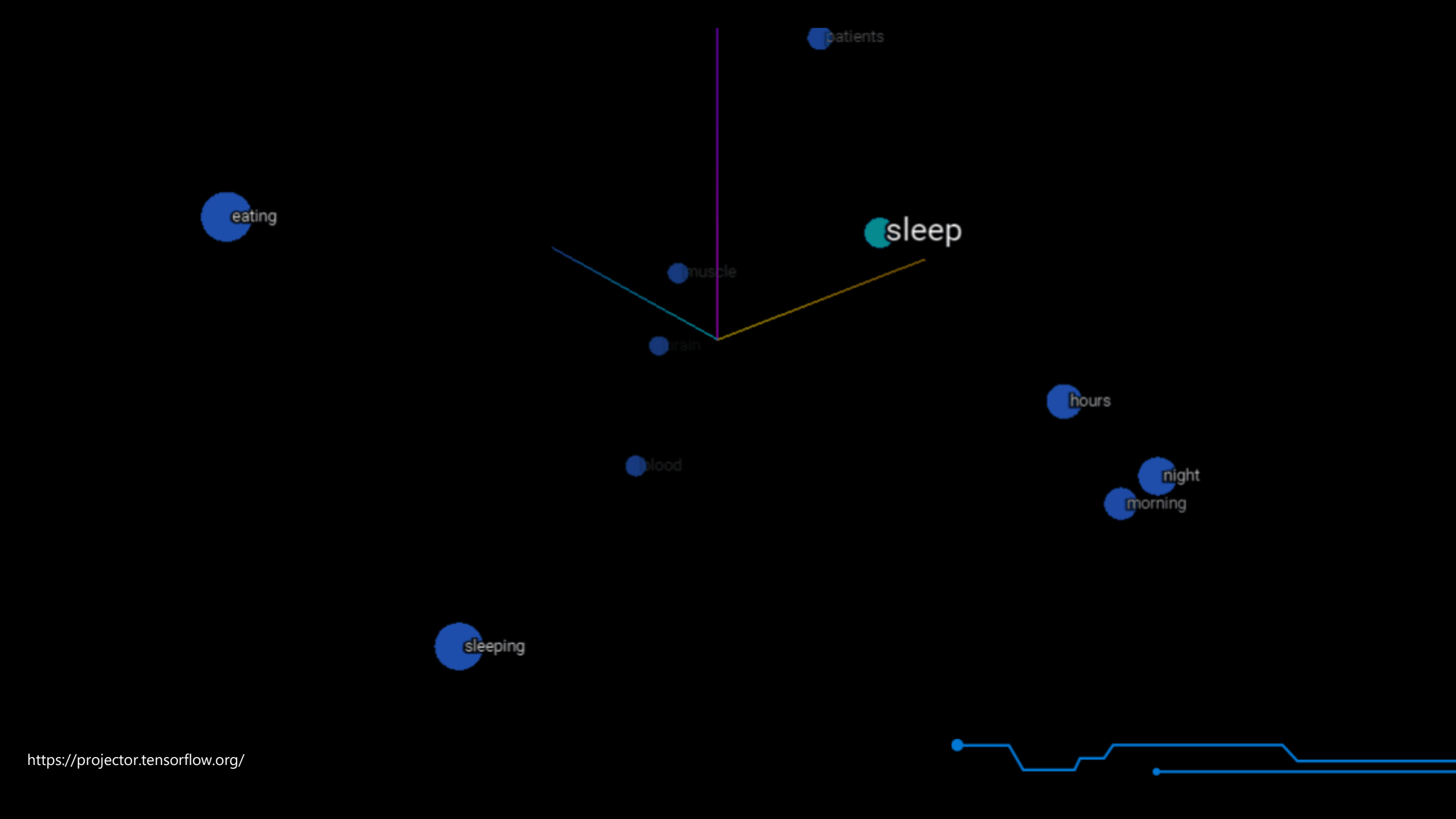
1) **Cosine Distance** – Text analysis, where direction is important
2) **Dot Product** – Evaluating similarity in terms of both direction and magnitude
3) **Euclidean Distance** – General purpose metric
4) **Manhattan Distance** – Grid-like structures, e.g., Image Processing.

2-Dimensional space

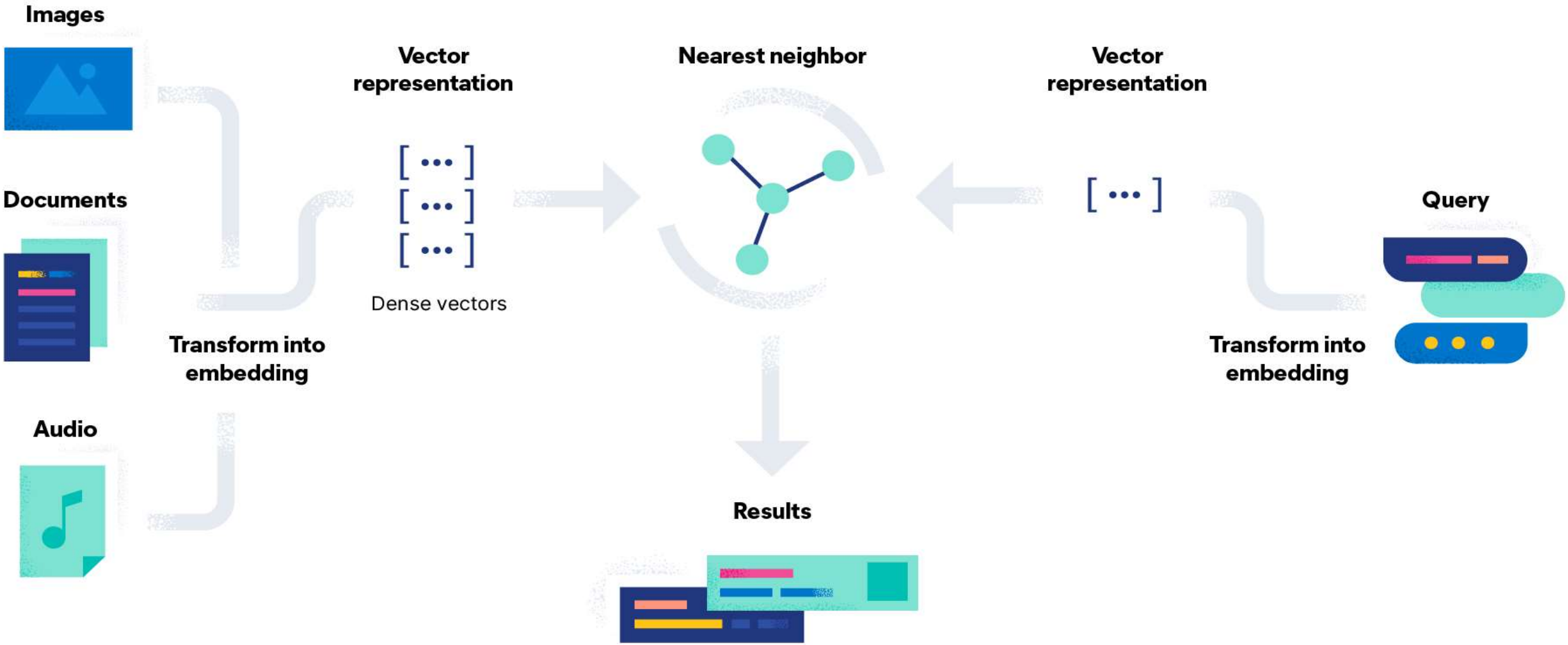




https://projector.tensorflow.org/



https://projector.tensorflow.org/



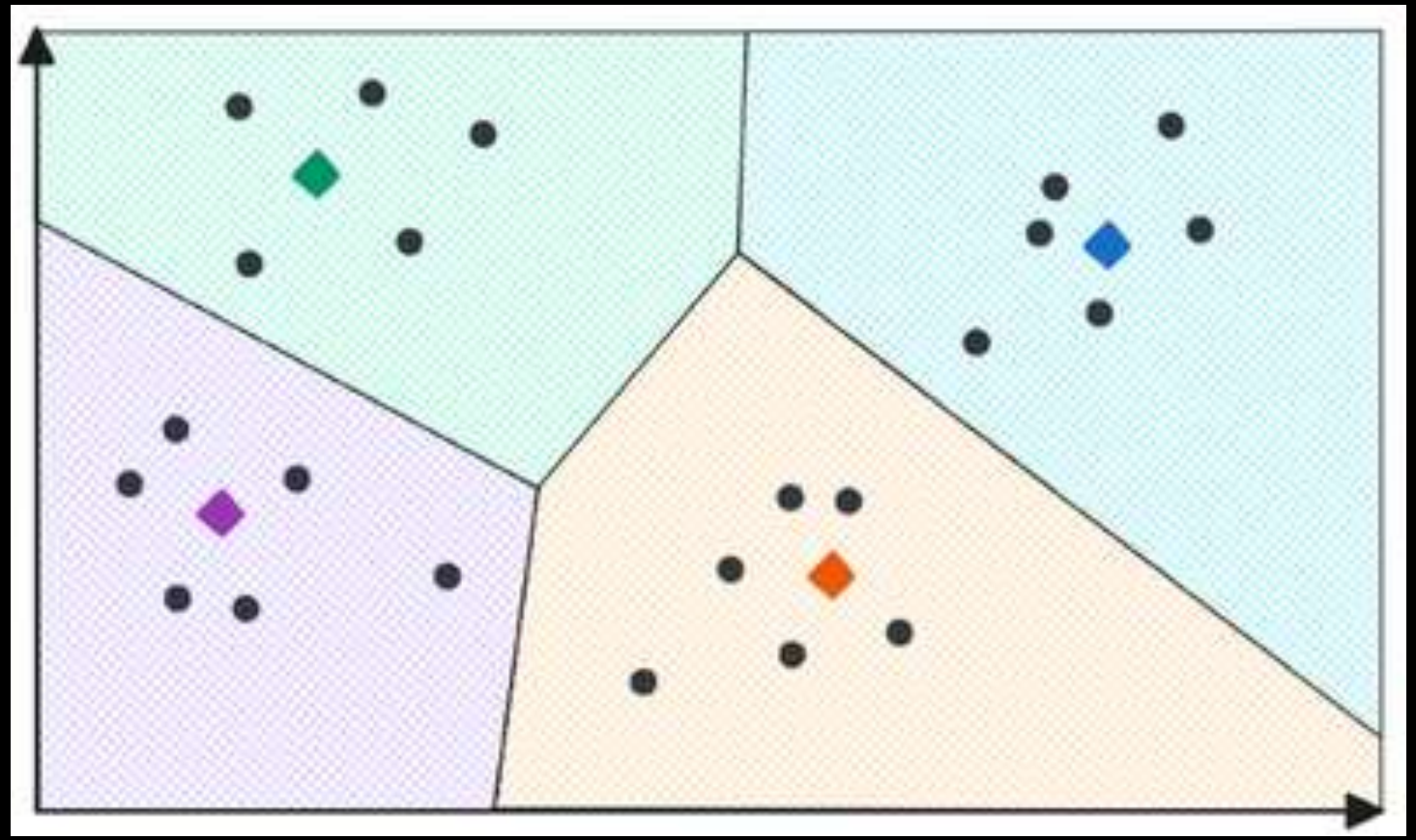
https://www.elastic.co/what-is/vector-database

Search algorithms

1) **Naive approach (KNN)** – Brute force search algorithm.
   - Measures the distance between the query and each vector
   - Sort all distances
   - Returns top k matches

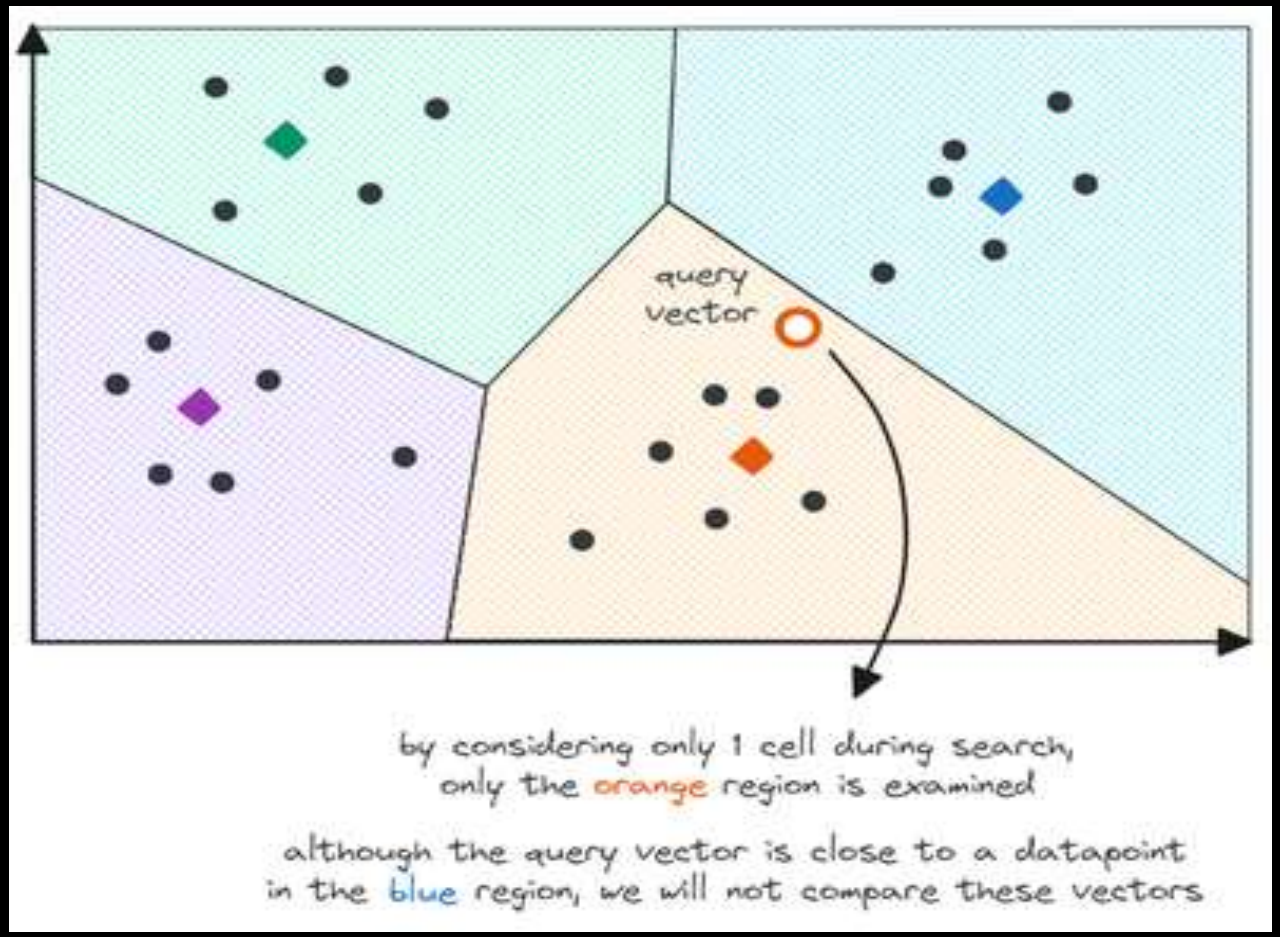
2) **Approximate Nearest Neighbors** – IVF or HNSW search algorithm.

IVFFlat Index



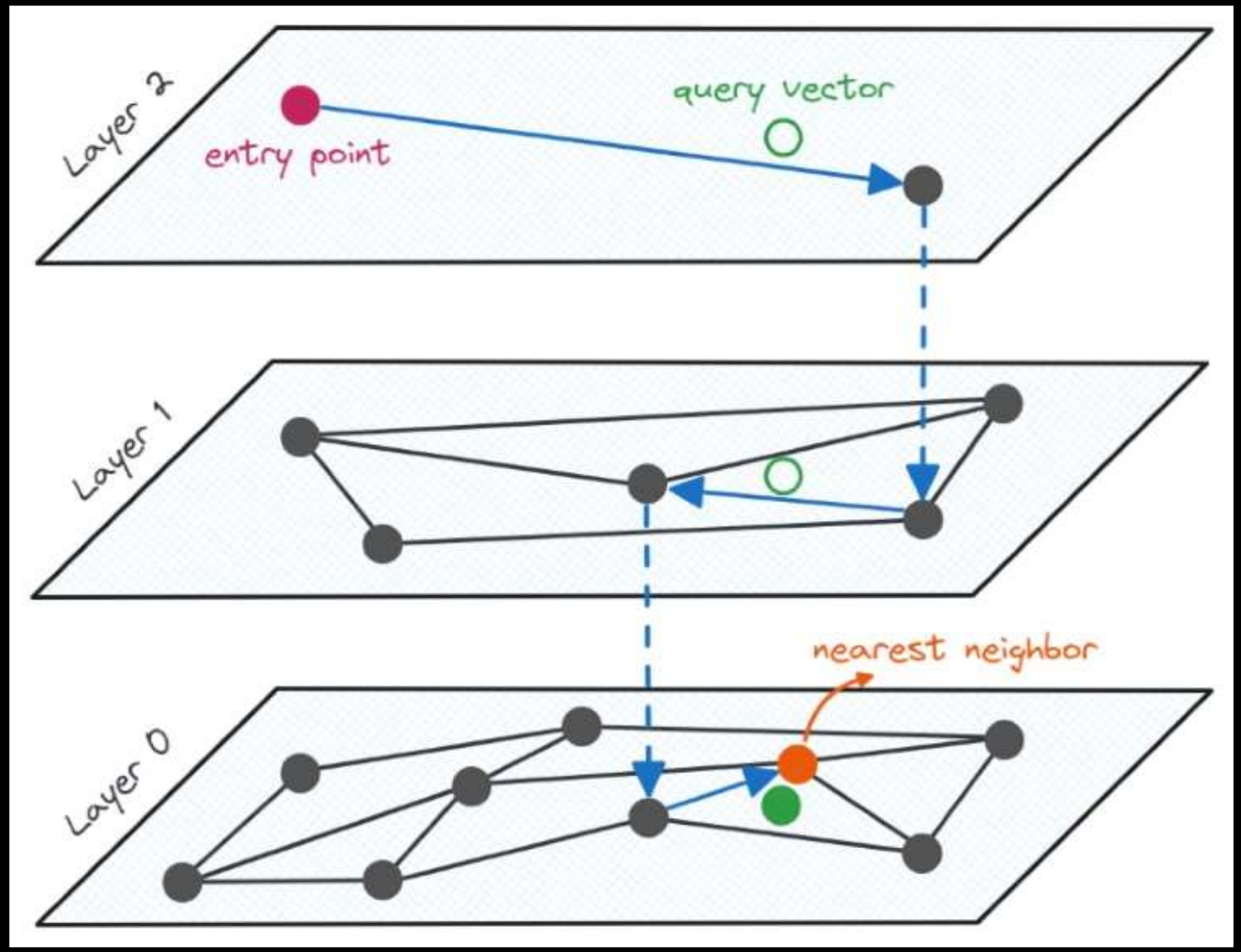
https://techcommunity.microsoft.com/t5/educator-developer-blog/speed-up-vector-search-in-azure-cosmos-db-for-postgresql-with/ba-p/4087077

IVFFlat Index



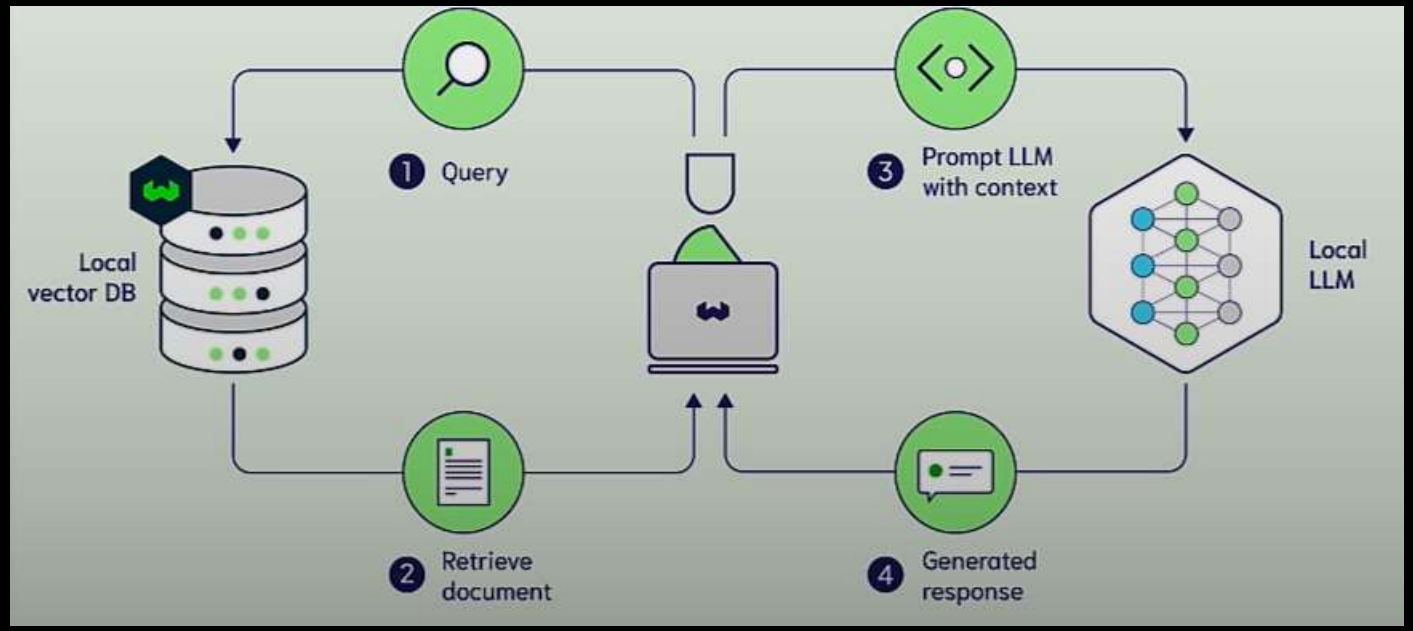
https://techcommunity.microsoft.com/t5/educator-developer-blog/speed-up-vector-search-in-azure-cosmos-db-for-postgresql-with/ba-p/4087077

HNSW Index

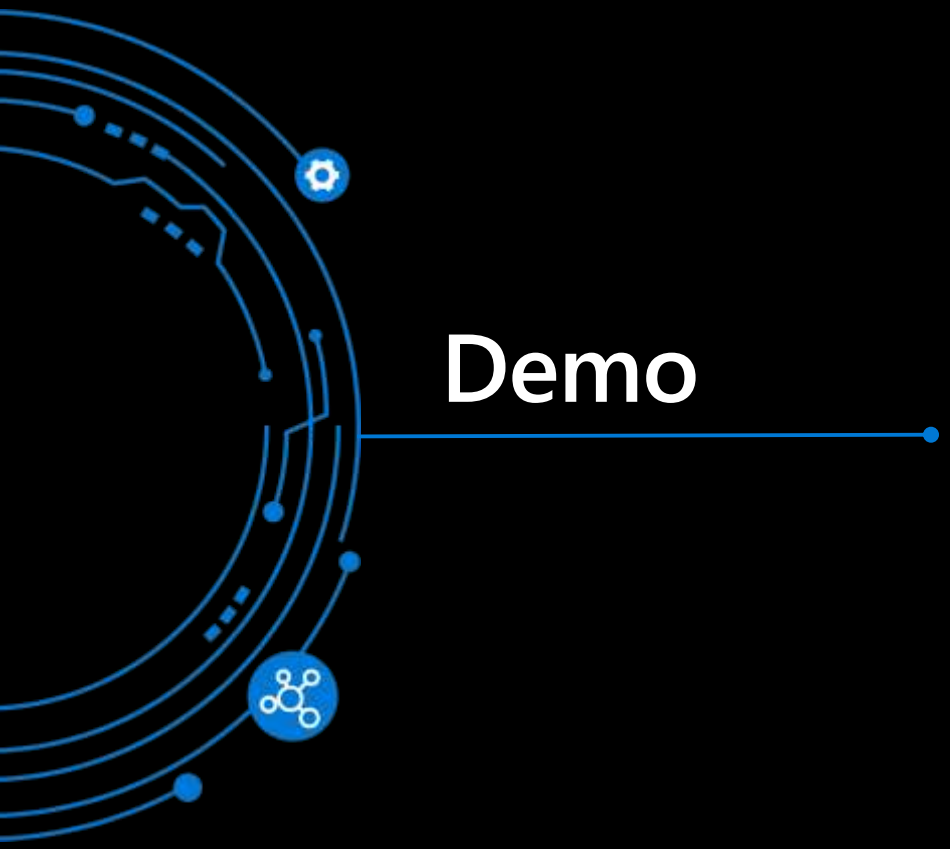


https://techcommunity.microsoft.com/t5/educator-developer-blog/speed-up-vector-search-in-azure-cosmos-db-for-postgresql-with/ba-p/4087077

RAG Worfklow



- Query a vector database
- Get relevant documents
- Combine the information to the prompt
- Send it to LLM to generate final response



Demo