# Building Generative AI Solutions with Vector Databases

## Vector search

La distancia entre dos puntos es un vector. Esta distancia se puede calcular de distintas formas: euclidiana, angular, KNN…

## Embeddings

* Embeddings are useful for a wide range of applications.
* For text, a common embedding model is named BERT. It starts by splitting the text into tokens, then we obtain embeddings for each token and finally average these into a sentence embedding.

Diagrama

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* For recommenders, matrix factorization is a popular way to generate embeddings.

Tabla

Descripción generada automáticamente

* For multimodal data like images, CLIP learns visual concepts via natural language and can generate embeddings for both text and images.

Diagrama

Descripción generada automáticamente

Texto

Descripción generada automáticamente

## Exercise 2

En este ejercicio hago un embedding de un texto de Ricky morty. Primero hago el embeding de todos los datos, luego hago una pregunta, le hago el embeding tmb y obtengo la respuesta.

## Vector Databases

There are three main types of tools for vector search:

* Vector index libraries.
* Traditional databases.
* Vector databases.

All three types of tools support the basic features of indexing and querying vectors.

Vector databases are more full-featured, more scalable, and more performant.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

## Exercise 3

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

## Advanced vector operations

Let's dig into ANN indices a bit to understand how vector databases speed up search on millions or even billions of vectors.

* ANN techniques are always a trade-off between speed and accuracy.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

* IVF and HNSW are popular ANN techniques to reduce the search space for a vector query.
* Quantization can be used to reduce the size of vectors that need to be stored, again at the expense of accuracy.

Escala de tiempo

Descripción generada automáticamente

## Exercise 4: Simple Recommender

Vector search is certainly critical for generative AI, but also has lots of other interesting applications as well. One very common one is building personalized recommendations. In this exercise, we'll take a small diversion and build a quick movie recommender using a vector database.