What is a 𝗩𝗲𝗰𝘁𝗼𝗿 𝗗𝗮𝘁𝗮𝗯𝗮𝘀𝗲?  
  
With the rise of Foundational Models, Vector Databases skyrocketed in popularity. The truth is that a Vector Database is also useful outside of a Large Language Model context.  
  
When it comes to Machine Learning, we often deal with Vector Embeddings. Vector Databases were created to perform specifically well when working with them:  
  
➡️ Storing.  
➡️ Updating.  
➡️ Retrieving.  
  
When we talk about retrieval, we refer to retrieving set of vectors that are most similar to a query in a form of a vector that is embedded in the same Latent space. This retrieval procedure is called Approximate Nearest Neighbour (ANN) search.  
  
A query here could be in a form of an object like an image for which we would like to find similar images. Or it could be a question for which we want to retrieve relevant context that could later be transformed into an answer via a LLM.  
  
Let’s look into how one would interact with a Vector Database:  
  
𝗪𝗿𝗶𝘁𝗶𝗻𝗴/𝗨𝗽𝗱𝗮𝘁𝗶𝗻𝗴 𝗗𝗮𝘁𝗮.  
  
1. Choose a ML model to be used to generate Vector Embeddings.  
2. Embed any type of information: text, images, audio, tabular. Choice of ML model used for embedding will depend on the type of data.  
3. Get a Vector representation of your data by running it through the Embedding Model.  
4. Store additional metadata together with the Vector Embedding. This data would later be used to pre-filter or post-filter ANN search results.  
5. Vector DB indexes Vector Embedding and metadata separately. There are multiple methods that can be used for creating vector indexes, some of them: Random Projection, Product Quantization, Locality-sensitive Hashing.  
6. Vector data is stored together with indexes for Vector Embeddings and metadata connected to the Embedded objects.  
  
𝗥𝗲𝗮𝗱𝗶𝗻𝗴 𝗗𝗮𝘁𝗮.  
  
7. A query to be executed against a Vector Database will usually consist of two parts:  
  
➡️ Data that will be used for ANN search. e.g. an image for which you want to find similar ones.  
➡️ Metadata query to exclude Vectors that hold specific qualities known beforehand. E.g. given that you are looking for similar images of apartments - exclude apartments in a specific location.  
  
8. You execute Metadata Query against the metadata index. It could be done before or after the ANN search procedure.  
9. You embed the data into the Latent space with the same model that was used for writing the data to the Vector DB.  
10. ANN search procedure is applied and a set of Vector embeddings are retrieved. Popular similarity measures for ANN search include: Cosine Similarity, Euclidean Distance, Dot Product.  
  
Some popular Vector Databases: Qdrant, Pinecone, Weviate, Milvus, Faiss, Vespa.

