A Speedy Introduction To Vector Databases

Steve Pousty
@thesteve0
VMWare Principal Dev Advocate

Agenda

1. Introduction to Vector Databases
2. What is different than RDBMs
3. Where to use them and what that means for you
4. Make you the life of the party

Oh great, another DB with vectors

What is a vector database

Easy answer - a data store that works with vectors

Let’s talk about “vectors”, aka embeddings

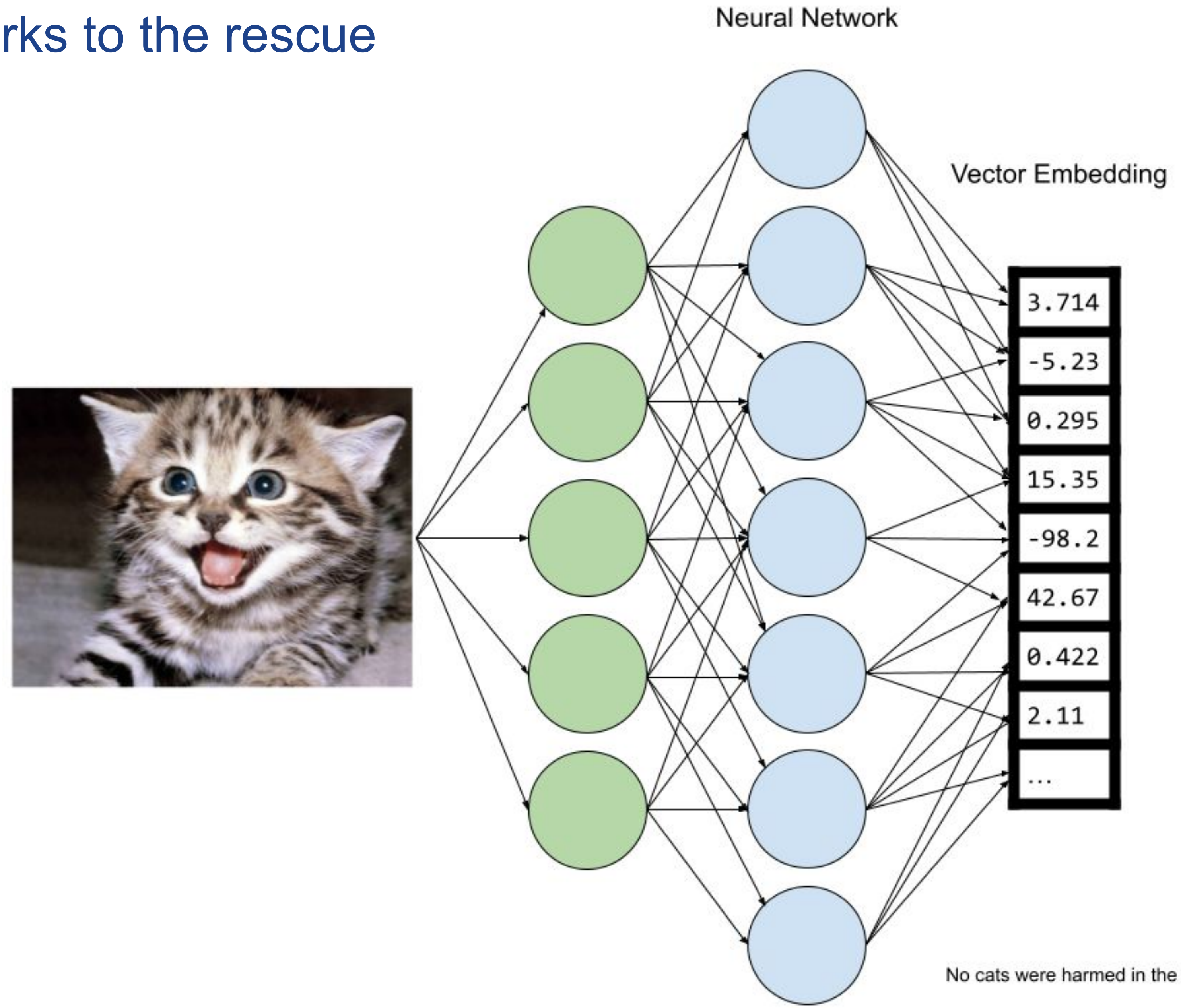
Turning Things into Numbers

Start with unstructured data - challenging for computers

2<3 ✅

❓

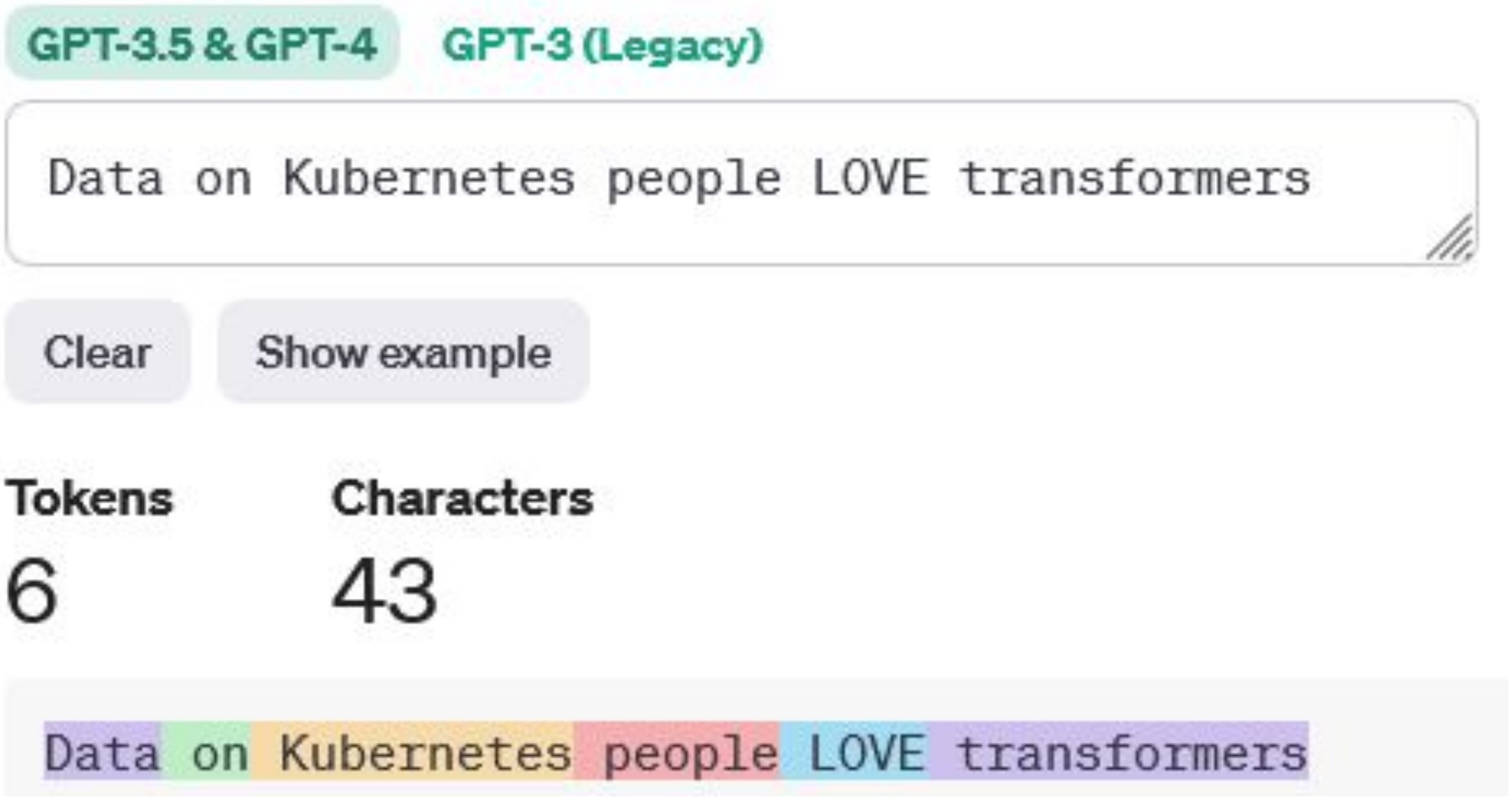
Neural Networks to the rescue



No cats were harmed in the making of this graphic

Brief Discussion on Tokens - NLP

API Costs and Context Length



Data on Kubernetes people LOVE transformers

Tokens
6

Characters
43

Data on Kubernetes people LOVE transformers

vmware®

Embeddings

There are more and more embedding models available to use.

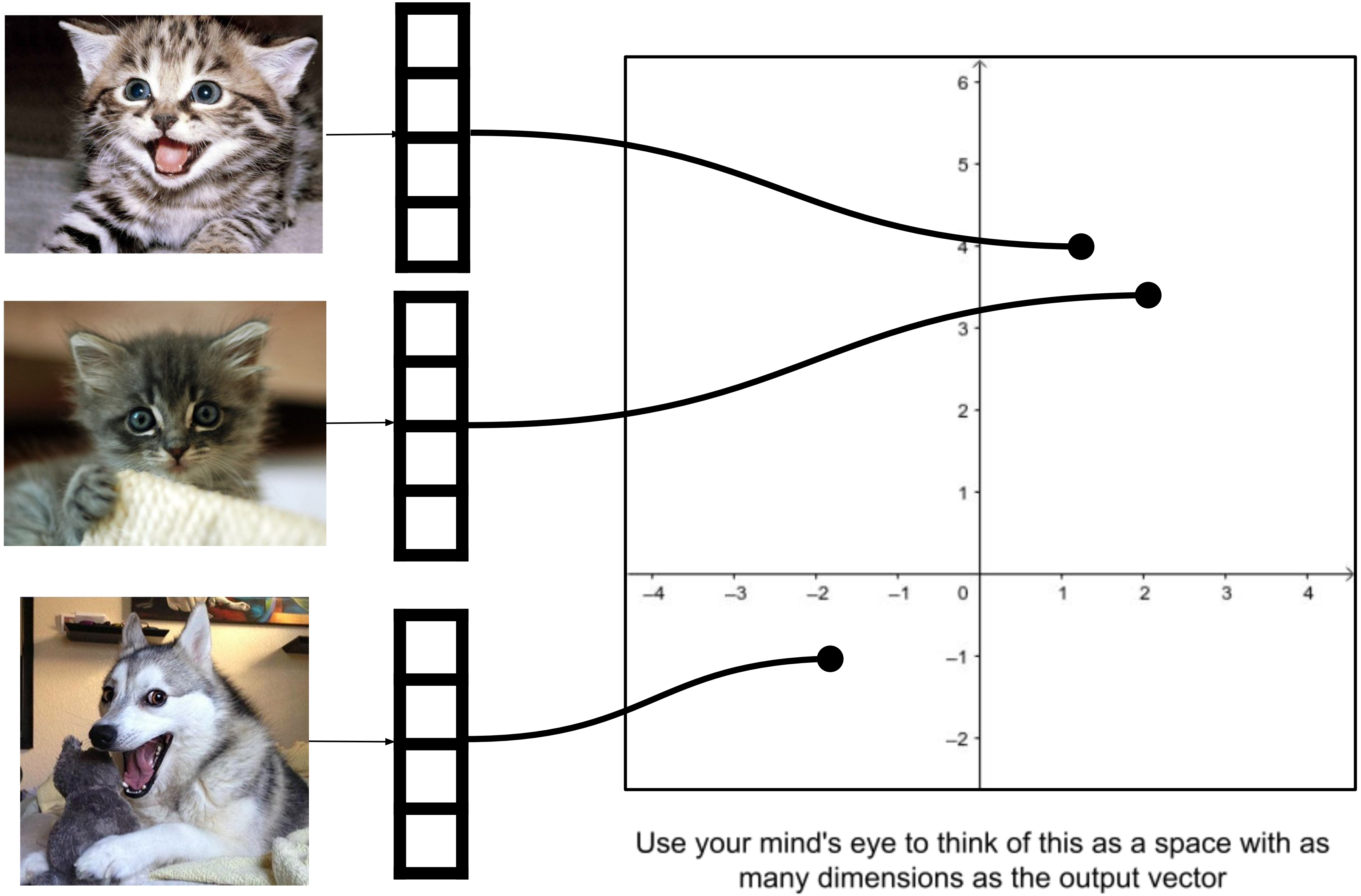
The ones we care about today are neural networks that have been pre-trained on large datasets.

There are several things to consider:

1. Appropriateness for task
2. Size of input
3. Length of output vector
4. Accuracy
5. Speed of computation

https://huggingface.co/models

Now into Vector Space

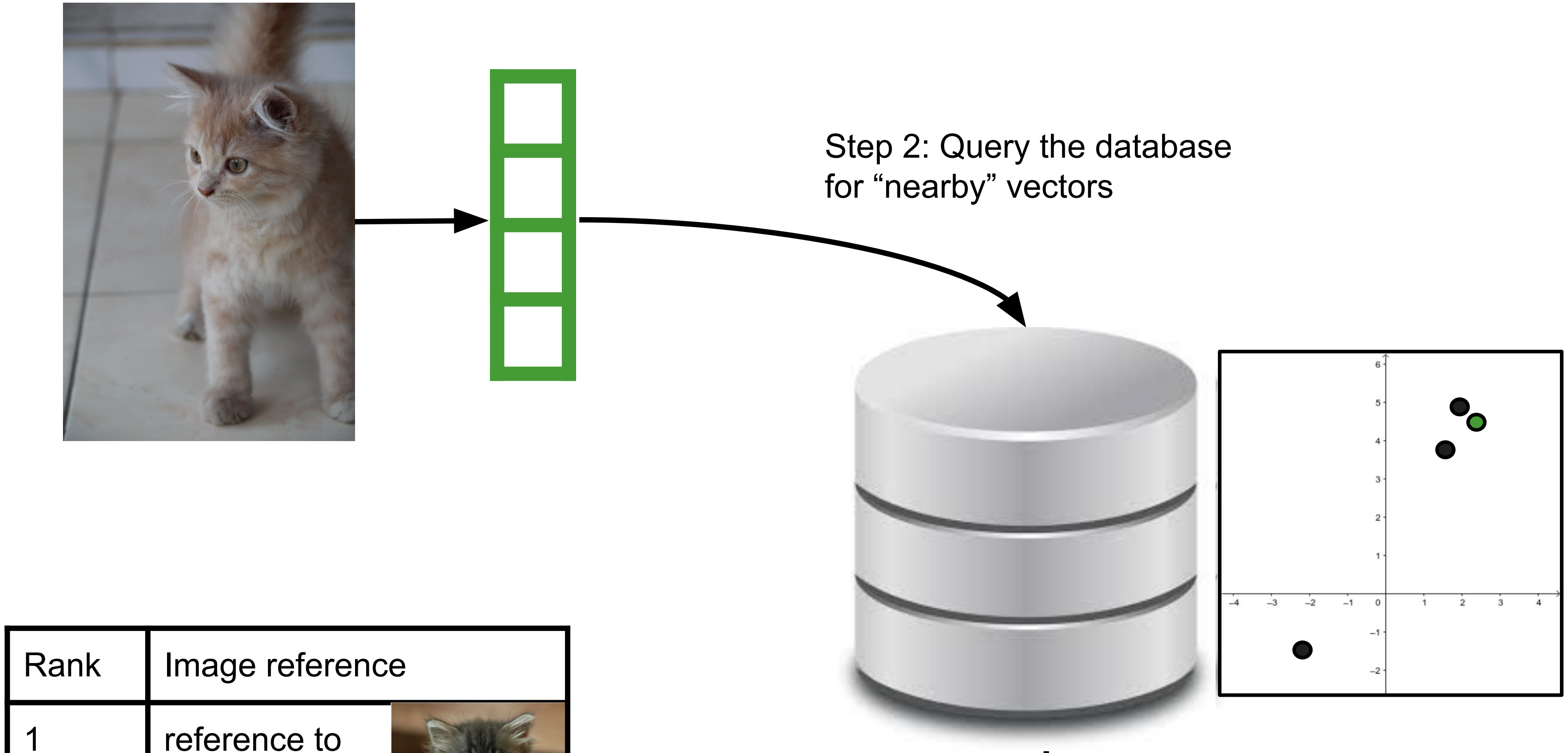


Use your mind's eye to think of this as a space with as many dimensions as the output vector

How to query

“What picture is similar to this picture”

Step 1: Cat to Vector



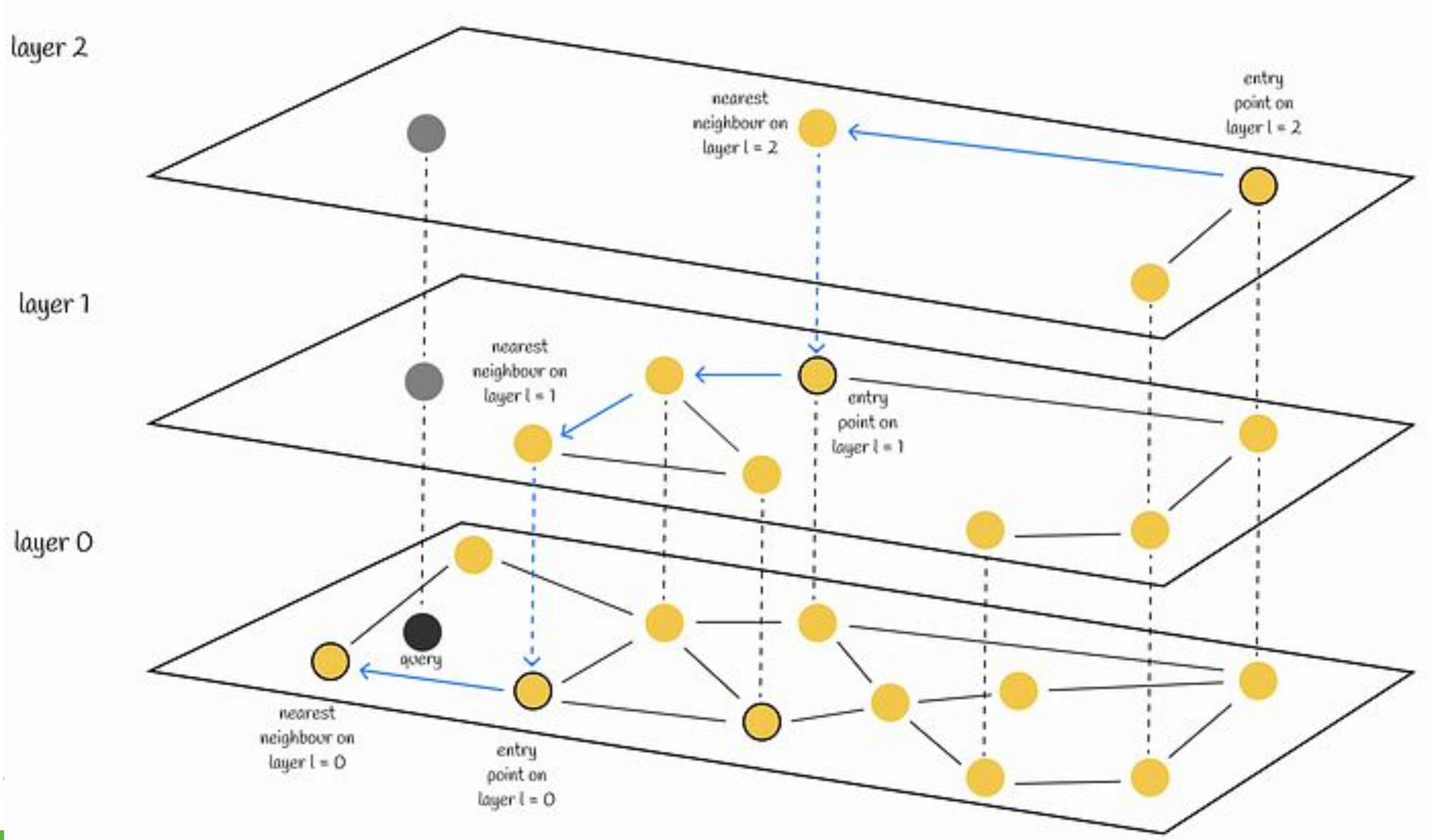
Step 2: Query the database for “nearby” vectors

Rank	Image reference
1	reference to
2	reference to
3	reference to

Step 3: Return results in decreasing distance order

Brief Discussion on HNSW

One of the most common Approximate Nearest Neighbor (ANN) indexing models


layer 2
nearest neighbour on layer l = 2
entry point on layer l = 2
layer 1
nearest neighbour on layer l = 1
entry point on layer l = 1
layer 0
query
nearest neighbour on layer l = 0
entry point on layer l = 0

What are they good for

Questions related to similarity

1. Not appropriate when exact search is the dominant use case
2. Specialized for a particular use case - they supplement your data infrastructure
3. Providing “memory” for your AI models
4. Reduce cost for running an AI infrastructure
5. **Interface between Data Science and Application Development**

Example use cases

1. Search (where results are ranked by relevance to a query vector)
2. Clustering (where items are grouped by similarity)
3. Recommendations (where related items are recommended)
4. Anomaly detection (where distant vectors little relatedness are identified)
5. Diversity measurement (where similarity distributions are analyzed)
6. Classification (where items are classified by their most similar label)

A Popular Example

Retrieval Augmented Generation (RAG)

Background Assumptions

1. You have some sort of generative text model to answer users’ questions.
2. OpenAI has trained their generative model on a broad corpus of texts
3. You have vectors for your documentation in a vector DB

The New Flow

4. User query -> embedding
5. Search you documentation with this embedding
6. Get back n closest documents
7. Add those documents as context (augmentation) to the original query
8. Send all the new text to OpenAI for prediction

Two types of Architecture

1. Add ons to existing databases - a new data type with new indices and functions.
2. Single purpose - not transactional like an RDBMS. BASE rather than ACID

Add-ons tend towards the same scaling properties as the base system.

Single purpose tend to be new and built with horizontal scaling in mind

What this means for you

1. They tend to be horizontally sharded/distributed so plan accordingly
2. A LOT of random reads so IOPs really matter
3. HNSW indices are big and should be in RAM
4. Streaming/ingestion pipeline is going to handle the embeddings
5. Reduce overall data stored in the DB - it’s a “compression” technique

6. **Given the newer bigger AI/ML push, they are definitely going to be part of your data infrastructure**

Sum it up

1. In ML/AI, vector refers to the generated numerical representation of unstructured data
2. The vector encodes “meaning” into a multidimensional space
3. Vector Databases allow you to store and query vectors
4. They handle questions related to similarity
5. They are usually distributed
6. Hang on, it should be an interesting ride

Thanks and Enjoy the Vectors!

Steve Pousty
@Thesteve0
https://bit.ly/dokvector