



Vector Search Best Practices

Yujian Tang

Upcoming Events



WEBINAR

Foundation Models are Going Multimodal

James Le
Developer Experience, Twelve Labs



July 27, 2023 | 9:00 AM PT



WEBINAR

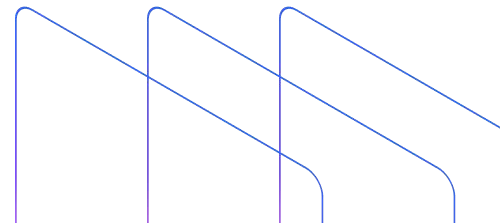
Designing Retrieval Augmentation for Generative Pipelines with Haystack

Tuana Çelik
Lead Developer Advocate, Deepset



August 10, 2023 | 9:00 AM PT

zilliz.com/event





Vector Search Best Practices

Yujian Tang

Speaker



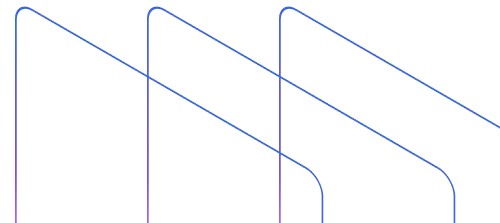
Yujian Tang

Developer Advocate, Zilliz

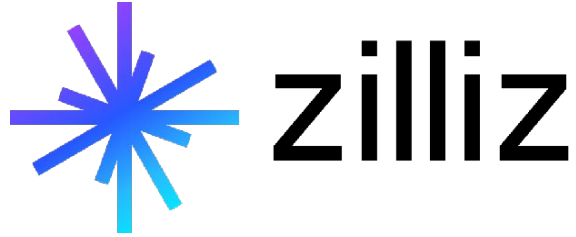
yujian@zilliz.com

<https://www.linkedin.com/in/yujiantang>

https://www.twitter.com/yujian_tang



Company



[@Zilliz_Universe](https://twitter.com/Zilliz_Universe)



linkedin.com/in/zilliz

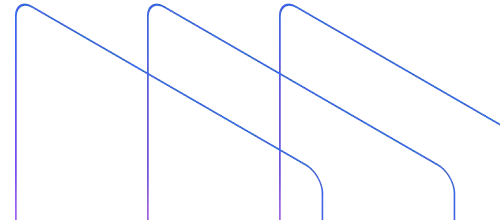


milvusio.slack.com

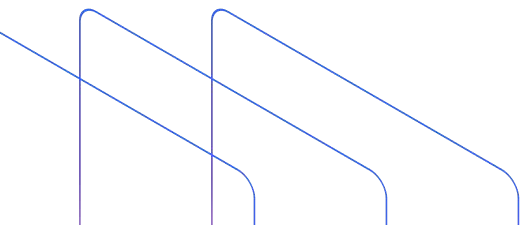


github.com/milvus-io/milvus

zilliz.com

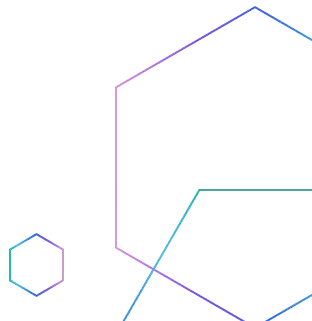
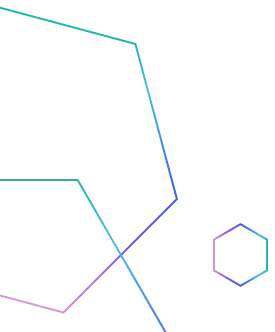


- 01 Why do I Need Vector Embeddings?**
- 02 What's a Vector Embedding?**
- 03 How do Vector Databases Work?**
- 04 Getting Started with a Vector Database**



01

Why do I Need Vector Embeddings?



Unstructured Data is Everywhere

Unstructured data is any data that does not conform to a predefined data model.

By 2025, IDC estimates there will be 175 zettabytes of data globally (that's 175 with 21 zeros), with **80% of that data being unstructured**. Currently, 90% of unstructured data is never analyzed.



Text



Images



Videos



and more!

Where can Vector Databases Help?



Image similarity search

Images made searchable and instantaneously return the most similar images from a massive database.



Molecular similarity search

Blazing fast similarity search, substructure search, or superstructure search for a specified molecule.



Question answering system

Interactive digital QA chatbot that automatically answers user questions.



Video similarity search

Converting key frames into vectors and storing the results, lets billions of videos be recommended in near real time.



Text search engine

Help users find the information they are looking for by comparing keywords against a database of texts.



Recommender system

Recommend information or products based on user behaviors and needs.



Audio similarity search

Quickly query massive volumes of audio data such as speech, music, sound effects, and surface similar sounds.



DNA sequence classification

Accurately sort out the classification of a gene in milliseconds by comparing similar DNA sequence.



Anomaly detection

Identifies data points, events, and/or observations that deviate from a dataset's normal behavior.

Adding Data to LLMs



Product Recommendations

Still Looking For Shoes?

Get 20% off your next order with **SHOES20OFF** coupon.



Brown Shoes

€78.00

Add to cart



Light Blue Shoes

€78.00

Add to cart



Black Shoes with Brown Details

€75.00

Add to cart



Brown and Black Shoes

€75.00

Add to cart

[Source](#)

Reverse Image Search



[Source](#)

Exploring Generative AI Use Cases



Image similarity search

Images made searchable and instantaneously return the most similar images from a massive database.



Video similarity search

Converting key frames into vectors and storing the results, lets billions of videos be recommended in near real time.



Audio similarity search

Quickly query massive volumes of audio data such as speech, music, sound effects, and surface similar sounds.



Molecular similarity search

Blazing fast similarity search, substructure search, or superstructure search for a specified molecule.



Text search engine

Help users find the information they are looking for by comparing keywords against a database of texts.



DNA sequence classification

Accurately sort out the classification of a gene in milliseconds by comparing similar DNA sequence.



Question answering system

Interactive digital QA chatbot that automatically answers user questions.



Recommender system

Recommend information or products based on user behaviors and needs.



Anomaly detection

Identifies data points, events, and/or observations that deviate from a dataset's normal behavior.

Exploring Generative AI Use Cases

Example

A company has 100,000s+ pages of proprietary documentation to enable their staff to service customers.

Problem

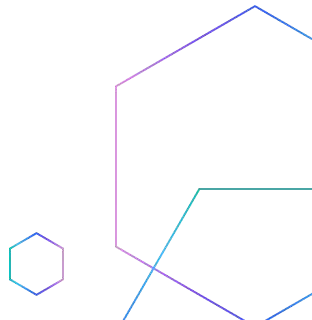
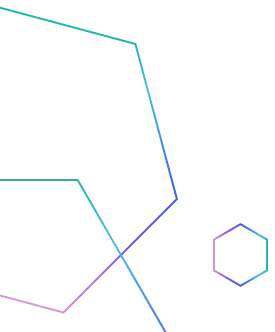
Searching can be slow, inefficient, or lack context.

Solution

Create internal chatbot with ChatGPT and a vector database enriched with company documentation to provide direction and support to employees and customers.

02

What's a Vector Embedding?



Introduction to Vectors

- Access to Domain Knowledge
- Semantic Search on Domain Knowledge via Vector Embeddings

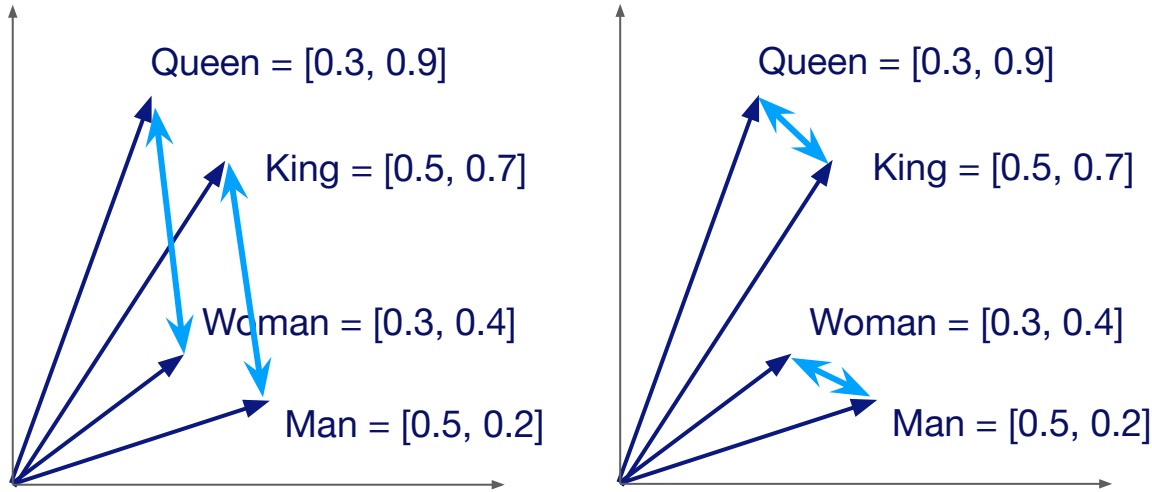
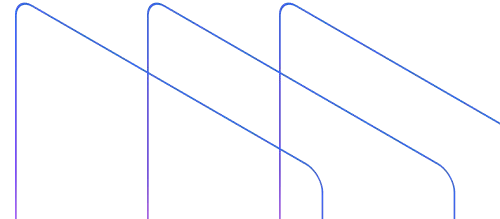
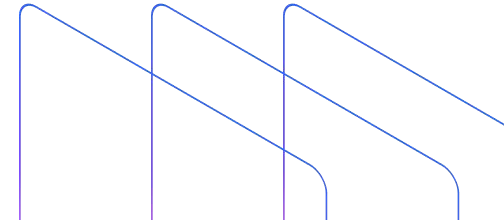
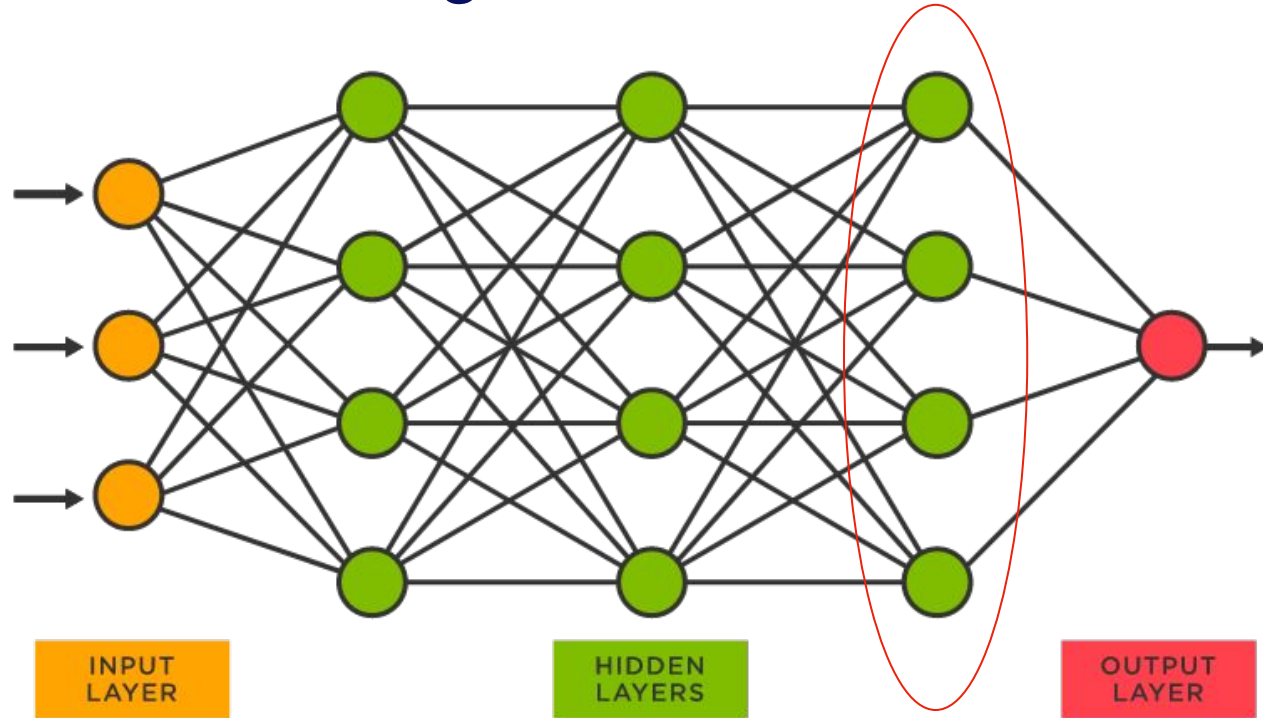


Image from [Sutor et al](#)



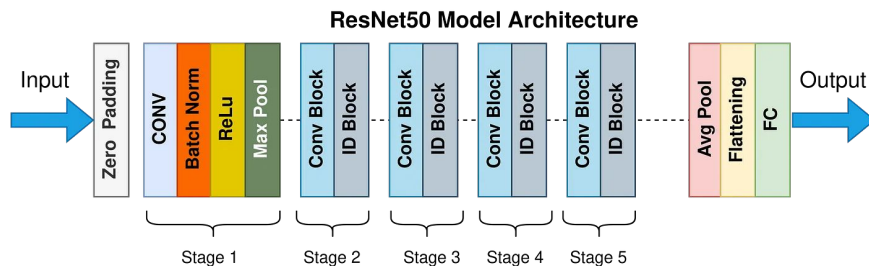
How are these generated?



Vector Embedding Model Choice



HUGGING FACE



What Do They Look Like?

Collection Details Schema Data Preview Vector Search

1-10 of 100 Datasets

```
"id": 125
"title_vector": [0.014838364,-0.017620698,0.039551493,0.015700748,-0.0011719975,-0.013021858,-0.010251275,0.01275...
"reading_time": "5"
"publication": "UX Collective"
"link": "https://uxdesign.cc/the-dawn-of-dark-mode-9636d1c9bcf0"
"responses": "0"
"title": "The dawn of Dark Mode"
"claps": "151"
```

↑ Hide 3 fields

🔍 Vector search

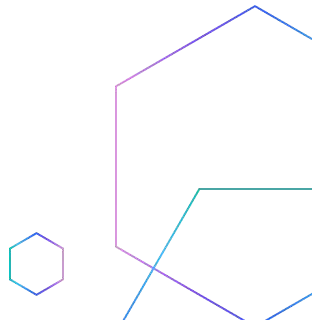
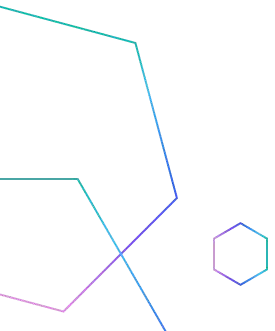
```
"id": 148
"title_vector": [0.034213345,-0.011796185,0.04076254,0.026783876,0.019659683,-0.025119903,0.016597062,0.024722276...
"reading_time": "4"
"publication": "The Startup"
"link": "https://medium.com/swlh/whats-next-for-neobanks-e304c900be98"
"responses": "0"
"title": "What's Next for NeoBanks?"
"claps": "136"
```

↑ Hide 3 fields

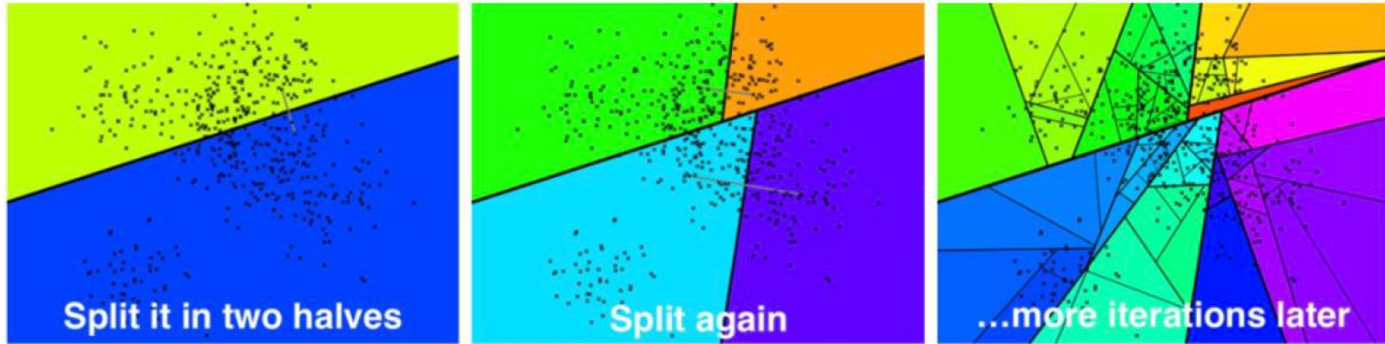
🔍 Vector search

03

How do Vector Databases Work?

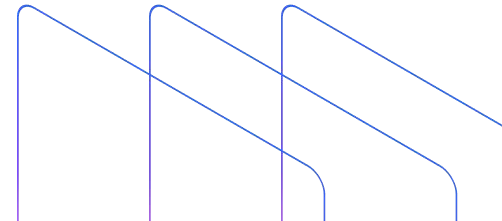


Approximate Nearest Neighbors Oh Yeah

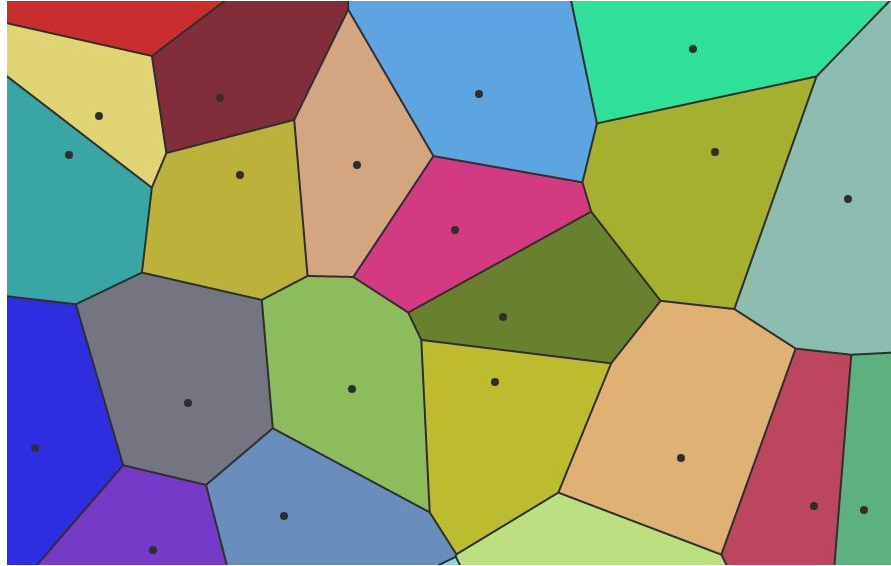


Source:

<https://sds-aau.github.io/M3Port19/portfolio/ann/>

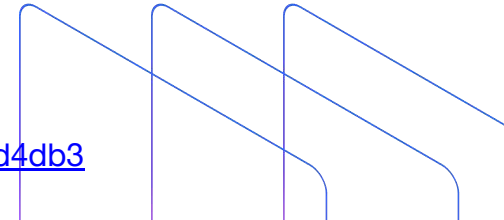


Inverted File Index

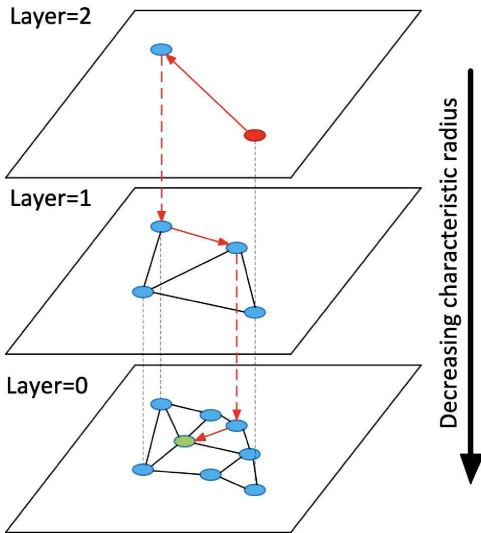


Source:

<https://towardsdatascience.com/similarity-search-with-ivfpq-9c6348fd4db3>

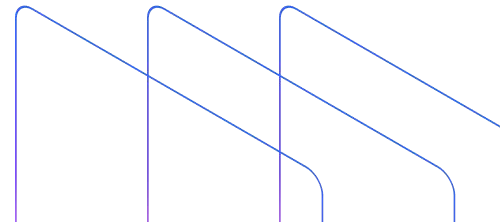


Hierarchical Navigable Small Worlds (HNSW)



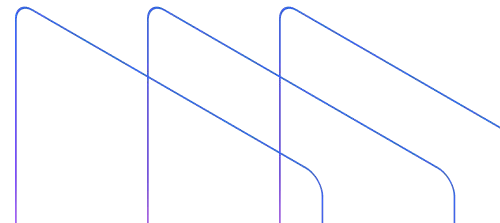
Source:

<https://arxiv.org/ftp/arxiv/papers/1603/1603.09320.pdf>



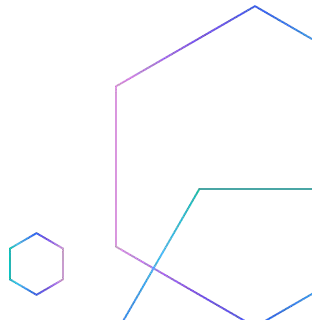
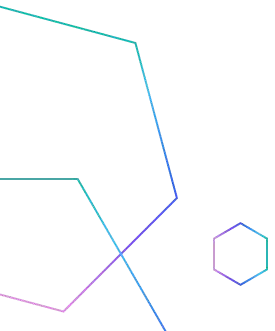
Why a Purpose-Built Vector Database?

- Vector search library
 - High-performance vector search
- Vector database
 - Advanced filtering (filtered vector search, chained filters)
 - Hybrid search (e.g. full text + dense vector)
 - Durability (any write in a db is durable, a library typically only supports snapshotting)
 - Replication / High Availability
 - Sharding
 - Aggregations or faceted search
 - Backups
 - Lifecycle management (CRUD, Batch delete, dropping whole indexes, reindexing)
 - Multi-tenancy
- How do I support different applications?
 - High query load
 - High insertion/deletion
 - Full precision/recall
 - Accelerator support (GPU, FPGA)
 - Billion-scale storage

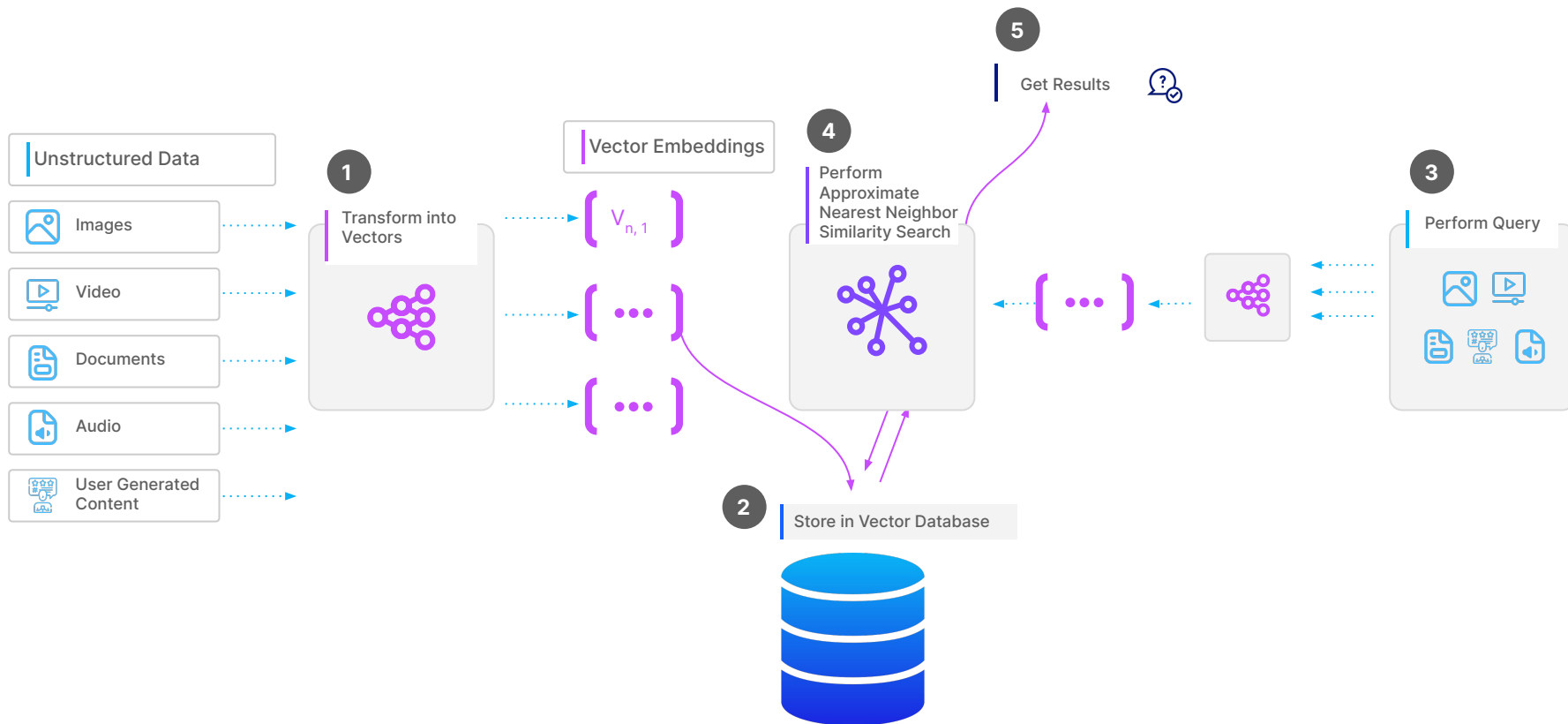


04

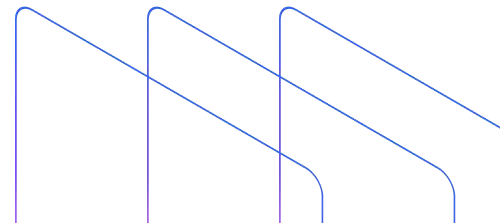
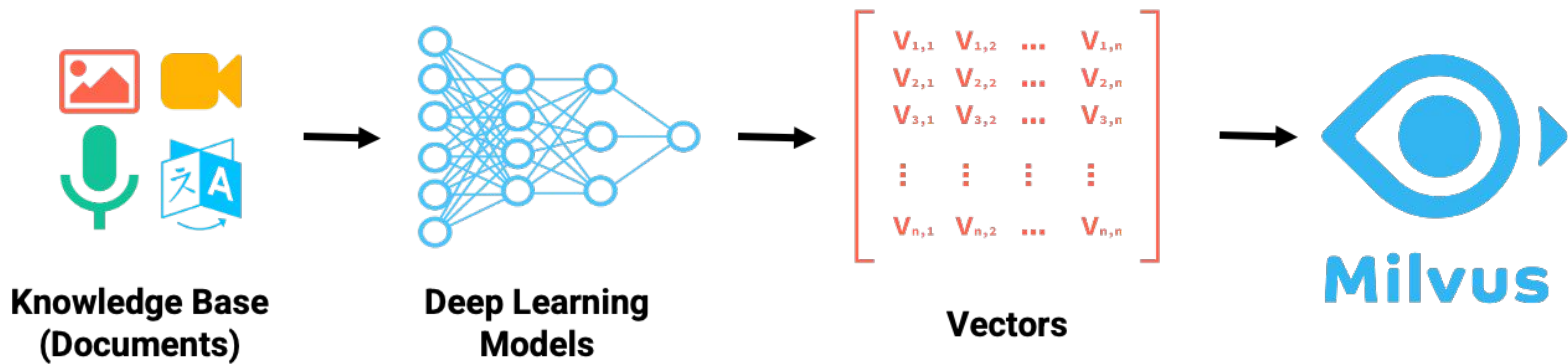
Getting Started with a Vector Database



How Similarity Search Works



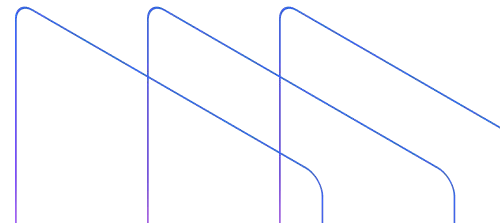
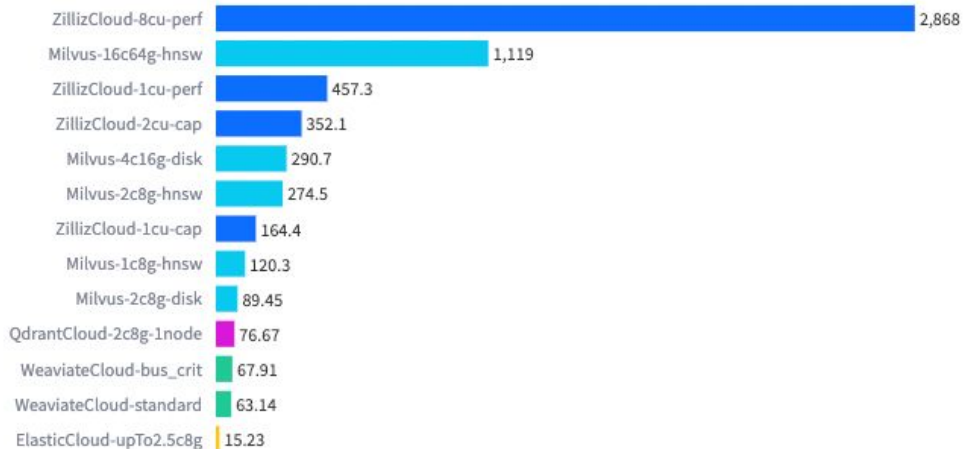
How Do We Implement This in Practice?



Vector Database Benchmarking

Search Performance Test (Medium Dataset)

Qps (more is better)





Get Started Today



github.com/milvus-io/milvus

