Database Management System

(Vector - Database)

Instructor: Thanh Binh Nguyen

February 1st, 2020

S3T
Smart Software System Team

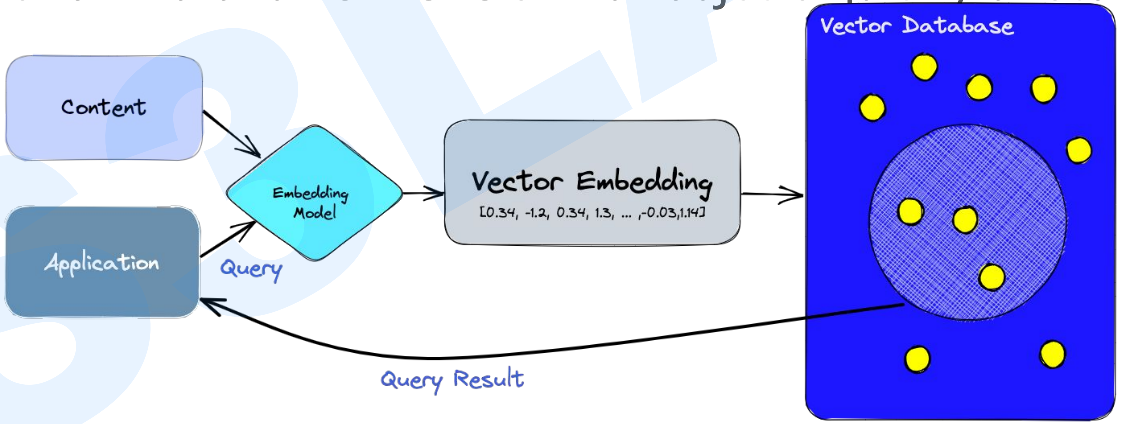
> “Data are becoming a new raw material of business.”
>
> – Craig Mundie, Microsoft

Vector Database

- **Indexes**, **stores**, and provides access to structured or unstructured data (e.g., text or images) alongside its **vector embeddings** (data's numerical representation).
- It allows users to **find** and **retrieve** similar objects quickly at scale in production.

Phase 1

CONTENT



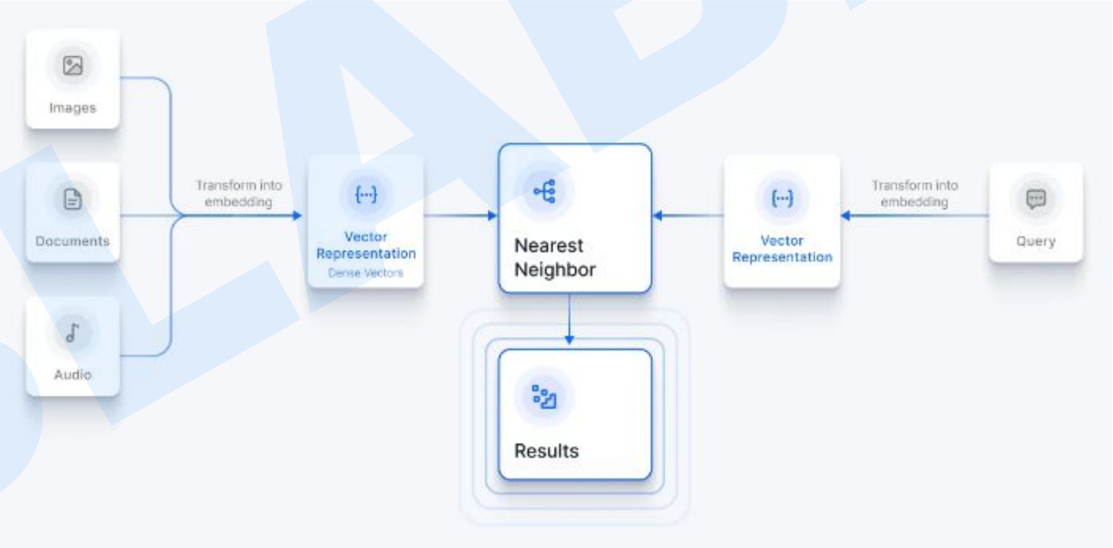
Phase 2

QUERY

When we use Vector DBMS?

Application

- Search engines
- Recommender systems
- Large Language Models
- Semantic search
- …



Vector Embeddings

- Structured Data: Neatly organized numbers in spreadsheets, easily be stored in tabular format.
- Unstructured Data: Images, text (e.g., documents, social media posts, or emails), or time series data (e.g., audio files, sensor data, or video).

  => Difficult to store it in an organized way, and find what you are looking for?

Vector Embeddings

- Numerically represent unstructured data without losing its semantic meaning in so-called **vector embeddings**. A vector embedding is just a long list of numbers, each describing a feature of the data object.
- **Vector embeddings** numerically capture the **semantic meaning** of the objects in relation to other objects. Thus, similar objects are grouped together in the vector space, which means the **closer two objects**, the **more similar** they are.

Vector Embeddings

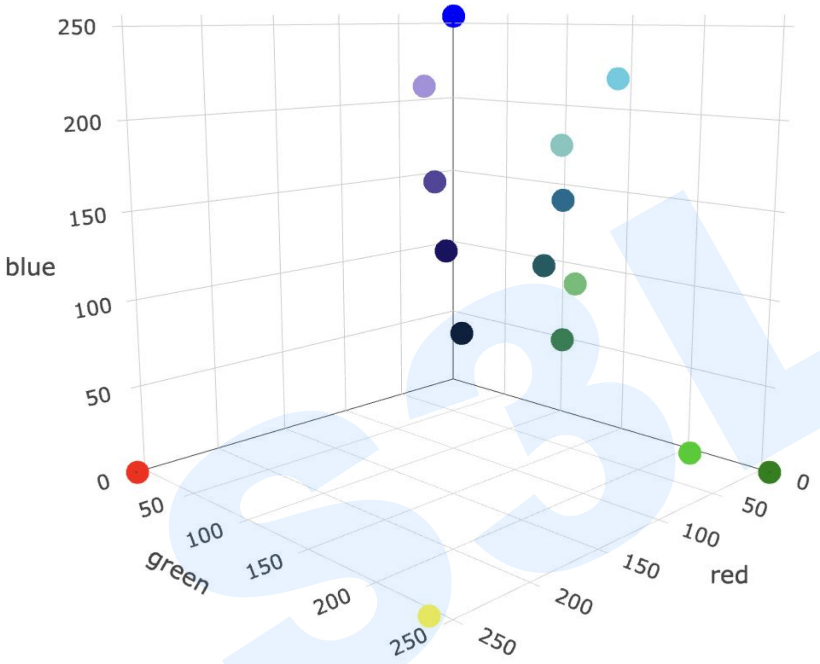
- Ex.

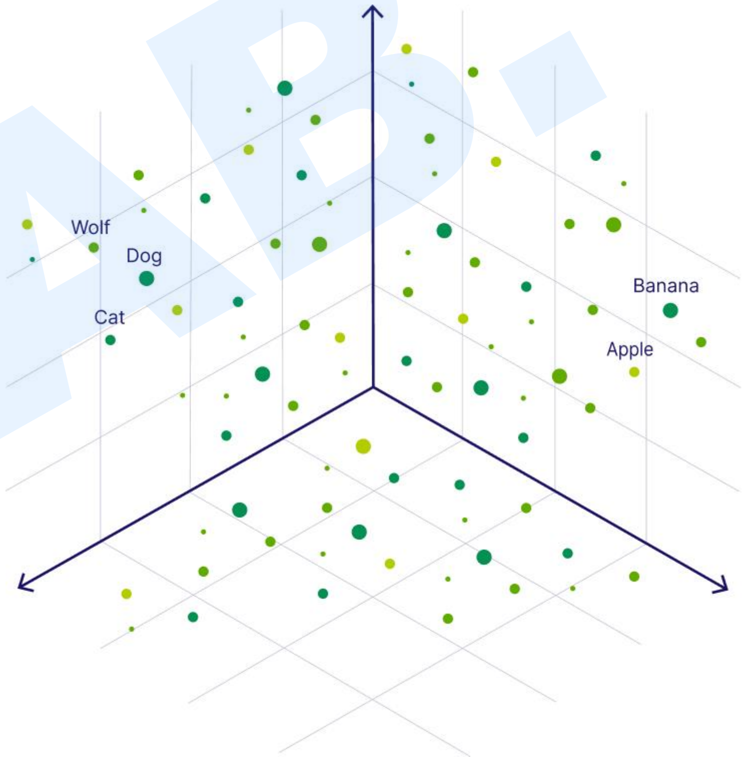
  RGB Color system: red, green, or blue -> [6, 205, 0] 🟩

- How about more complex data: such as words, sentences, or text ?

  Machine Learning models enable us to represent the contextual meaning of, e.g., a word as a vector because they have learned to represent the relationship between different words in a vector space. These types of Machine Learning models that can generate **embeddings** from **unstructured data** are also called **embedding model** or **vectorizer**.

Vector Embeddings

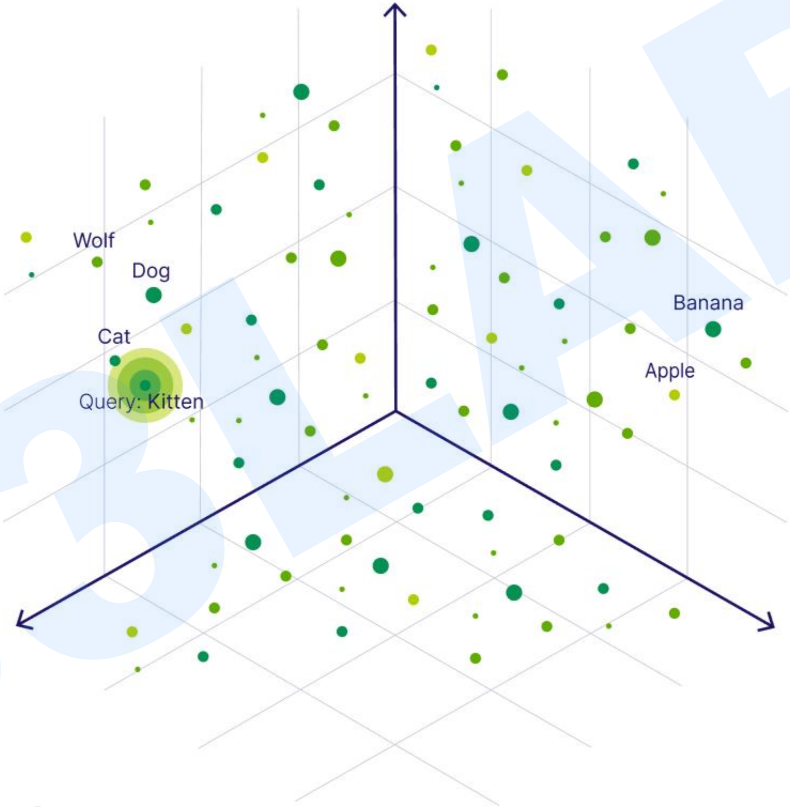




Vector Search

- **Vector embeddings** allow us to find and retrieve similar objects from the vector database by searching for objects that are close to each other in the vector space, which is called **vector search**, **similarity search**, or **semantic search**.
- For search, we can generate a vector embedding for the query term - also called a **Query Vector** - and retrieve all its nearest neighbors.

Vector Search



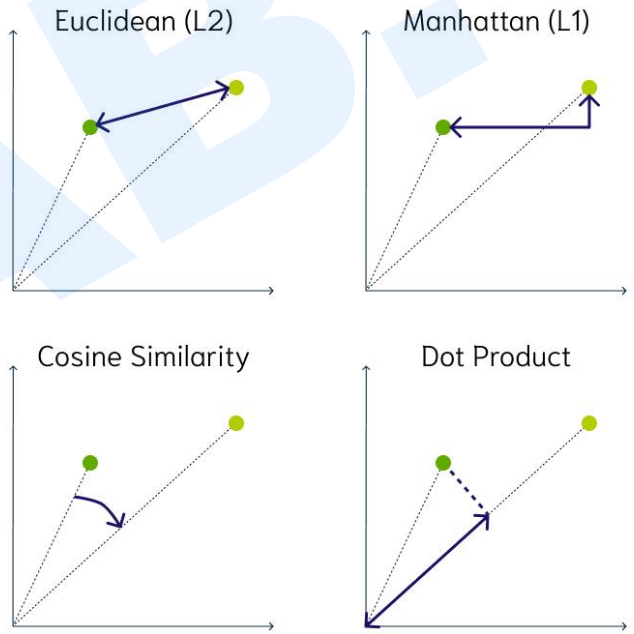
Vector Search

- As the concept of **semantic search** is based on the **contextual meaning**, it allows for a more human-like search experience by retrieving relevant search results that match the user's intent. This advantage makes vector search important for applications, that are e.g., sensitive to typos or synonyms.
- The numerical representation of a data object allows us to **apply mathematical operations** to them, such as calculating the **<u>distance</u>** between two vector representations to determine their **<u>similarity</u>**. To calculate the distance between two vectors, you can use several similarity measures.

Vector Search

Distance metrics

- Squared Euclidean or L2-squared distance
- Manhattan or L1 distance
- Cosine similarity
- Dot product
- Hamming distance

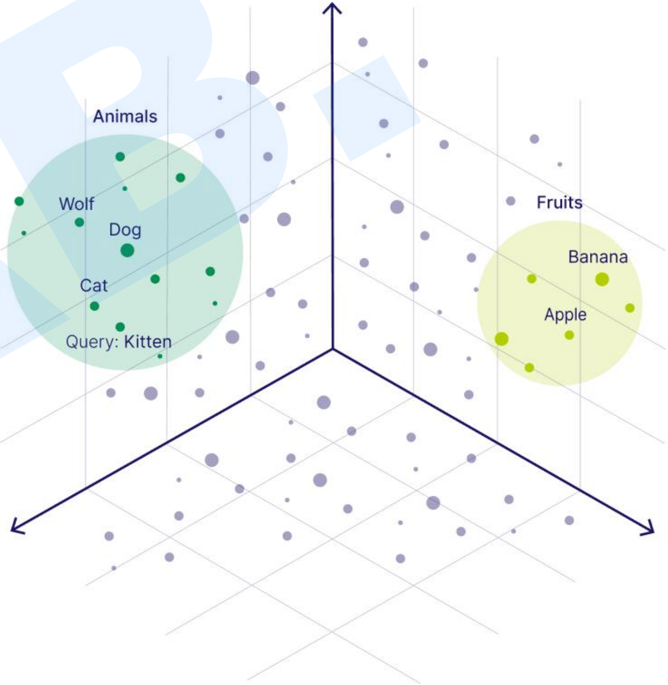


Vector Indexing

- The process of **organizing vector embeddings** in a way that data can be retrieved efficiently.
- Calculating the similarity between your query vector and every entry in the vector database requires a lot of computational resources, especially if you have large datasets with **millions** or even **billions of data points**, because the required calculations increase linearly (O(n)) with the dimensionality and the number of data points.
- Indexing enables **fast retrieval** at query time, but it can take **a lot of time to build the index initially**.

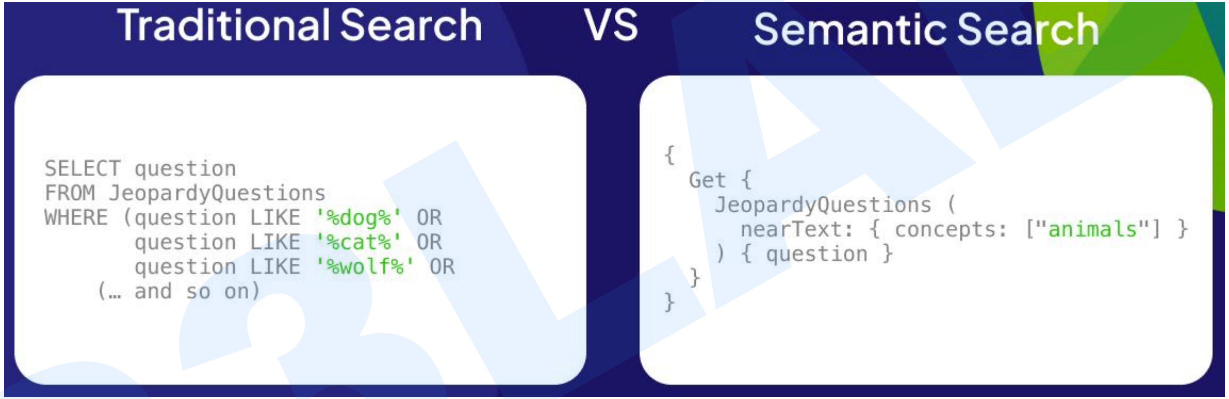
Vector Indexing

- How to store & search billions of embeddings?
- To find the closest items to a given query vector
  - k-Nearest Neighbors (kNN)
  - Approximate nearest neighbor (ANN)
    - Clustering-based index (e.g., FAISS)
    - Proximity graph-based index (e.g., HNSW)
    - Tree-based index (e.g., ANNOY)
    - Hash-based index (e.g., LSH)
    - Compression-based index (e.g., PQ or SCANN)



Tool Landscape around Vector Databases

- Vector Database vs. Traditional (Relational) Database



```
SELECT question
FROM JeopardyQuestions
WHERE (question LIKE '%dog%' OR
       question LIKE '%cat%' OR
       question LIKE '%wolf%' OR
    (… and so on)
```

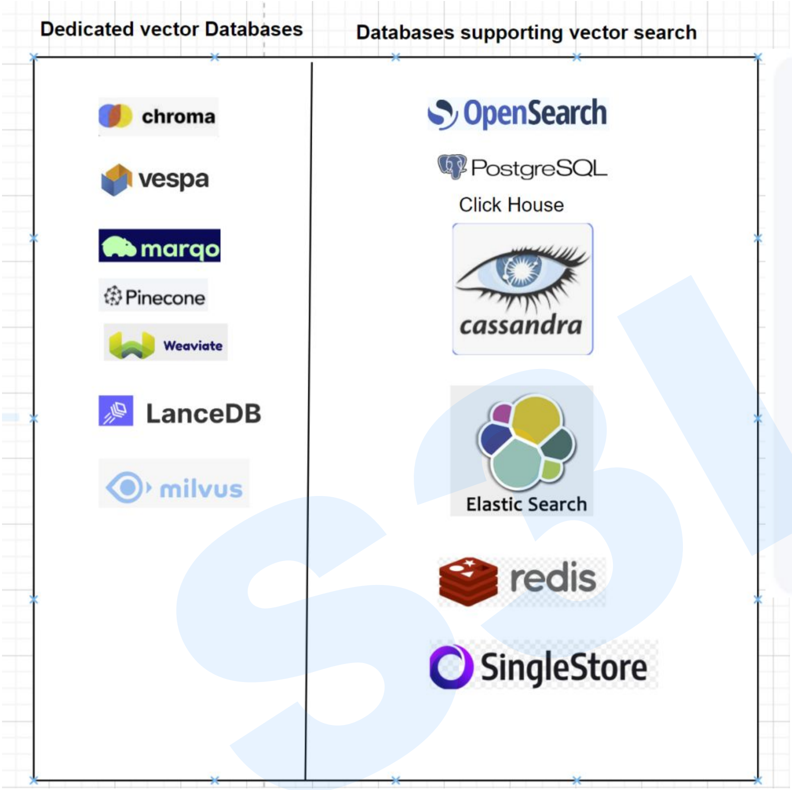
```
{
  Get {
    JeopardyQuestions (
      nearText: { concepts: ["animals"] }
    ) { question }
  }
}
```

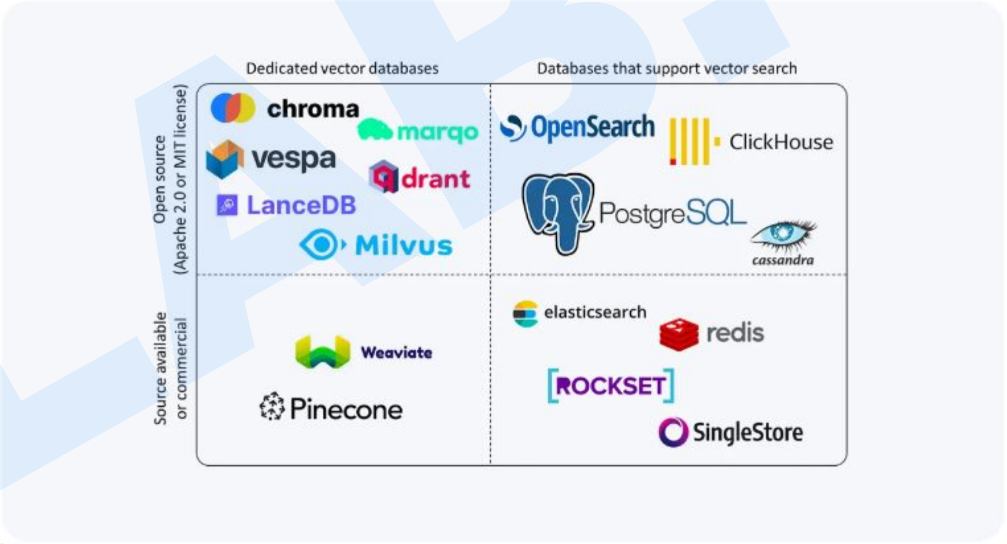
- Vector Database vs. Vector-Capable Database (SQL and NoSQL)
  - Vector-Capable DB usually don’t index the vector embeddings

Tool Landscape around Vector Databases

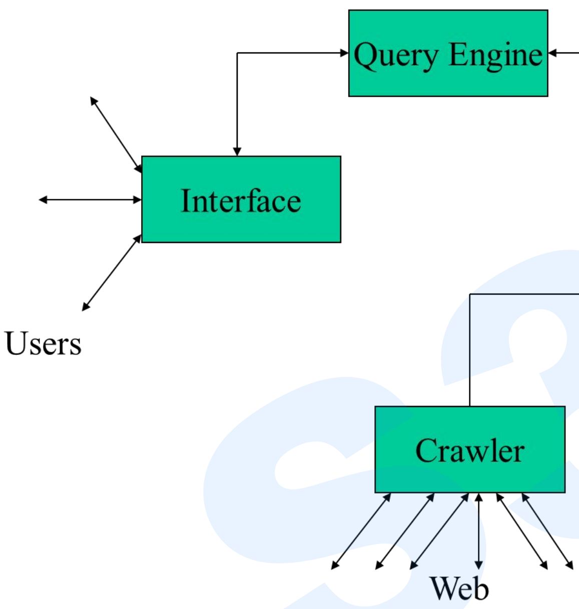
- Vector Database vs. Vector Indexing Library
  - Updatability: The index data is immutable, and thus, no real-time updates are possible.
  - Scalability: Most vector libraries cannot be queried while importing your data, which can be a scalability concern for applications that require importing millions or even billions of objects.

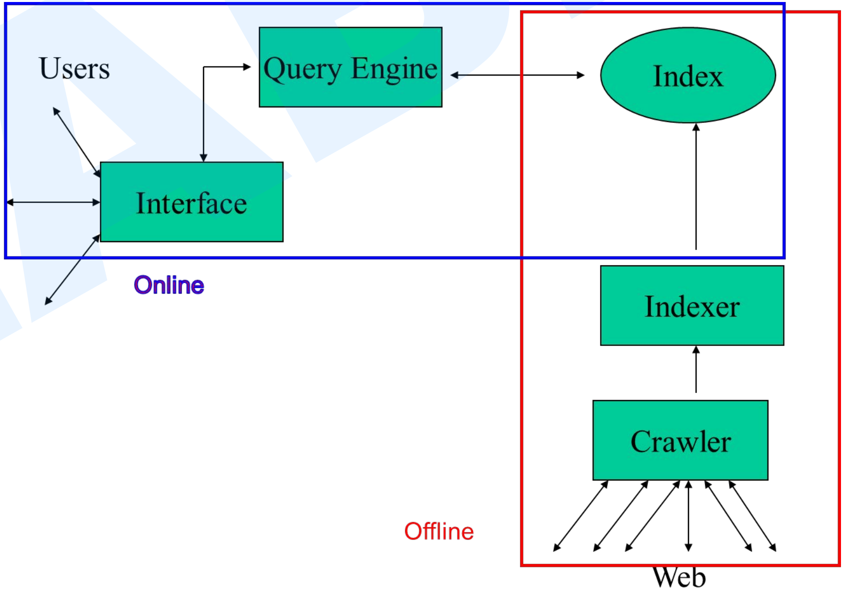
Tool Landscape around Vector Databases



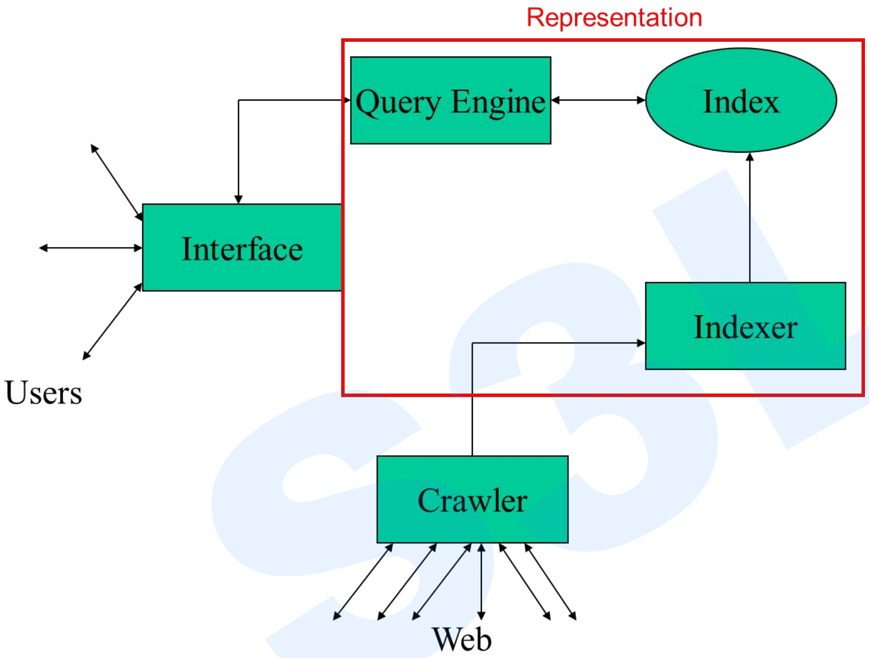

Dedicated vector databases
Databases that support vector search
Open source (Apache 2.0 or MIT license)
Source available or commercial
chroma
marqo
vespa
Qdrant
LanceDB
Milvus
OpenSearch
ClickHouse
PostgreSQL
cassandra
Weaviate
Pinecone
elasticsearch
redis
ROCKSET
SingleStore

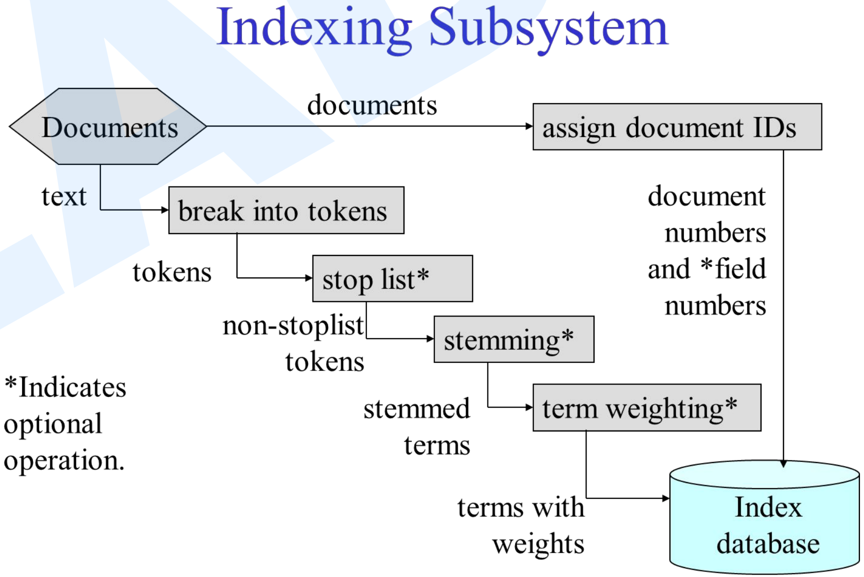
User cases: Web Search Engine





User cases: Web Search Engine

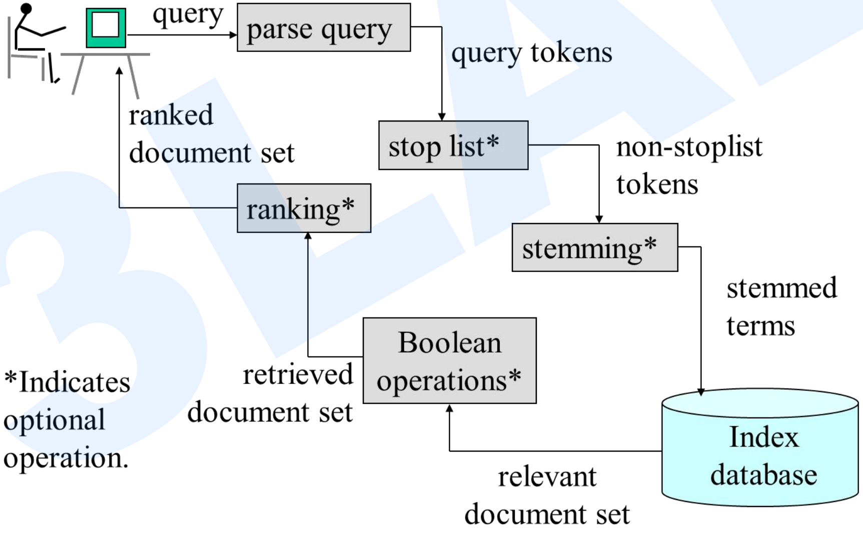




*Indicates optional operation.

User cases: Web Search Engine

Search Subsystem



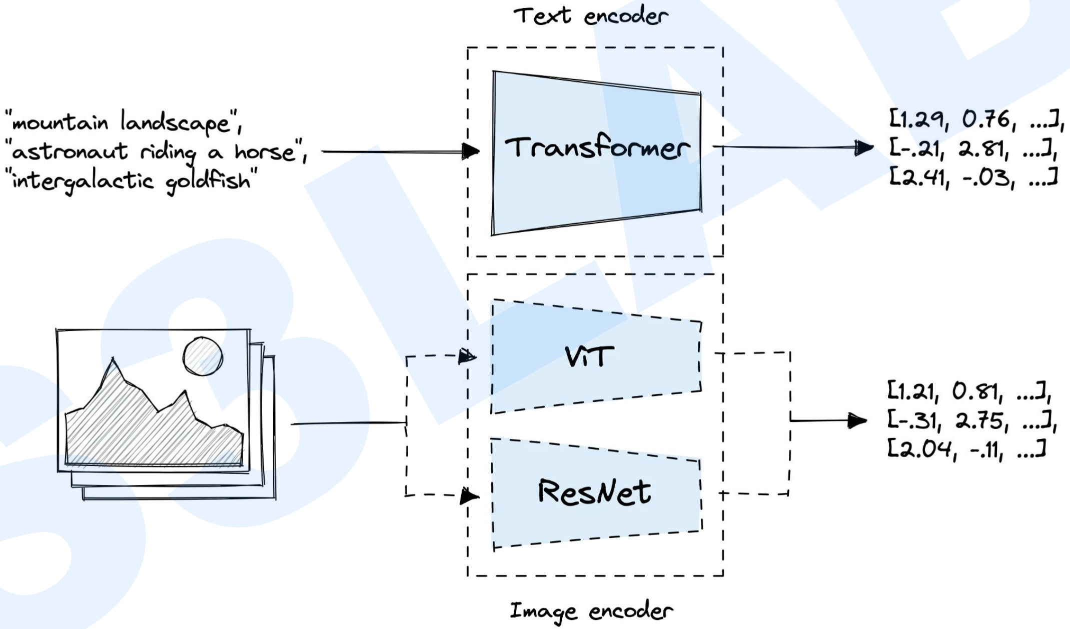
*Indicates optional operation.

User cases: Image Search With CLIP

- CLIP architecture consists of two main components:
  - The text encoder’s backbone is a transformer model [2], and the base size uses 63 millions-parameters, 12 layers, and a 512-wide model containing 8 attention heads.
  - The image encoder, on the other hand, uses both a Vision Transformer (ViT) and a ResNet50 as its backbone, responsible for generating the feature representation of the image.

User cases: Image Search With CLIP

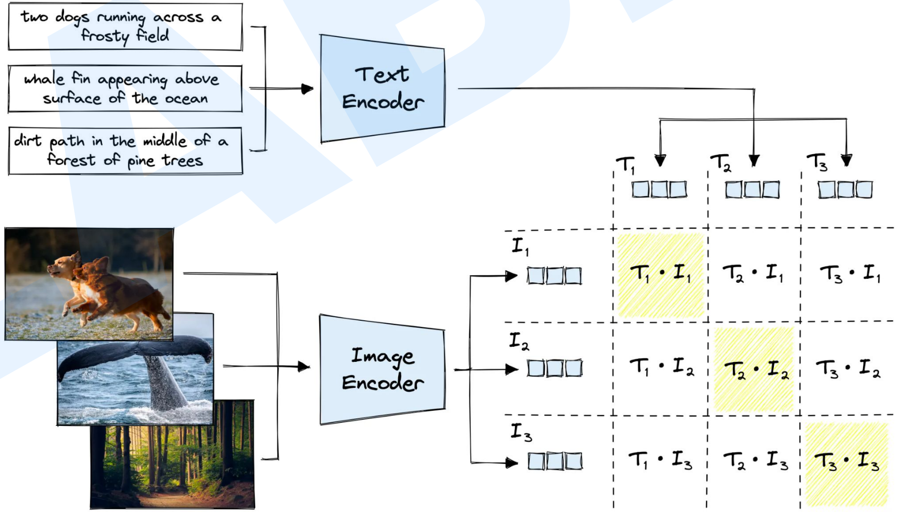
- CLIP architecture consists of two main components:



User cases: Image Search With CLIP

How it works - Contrastive pre-training

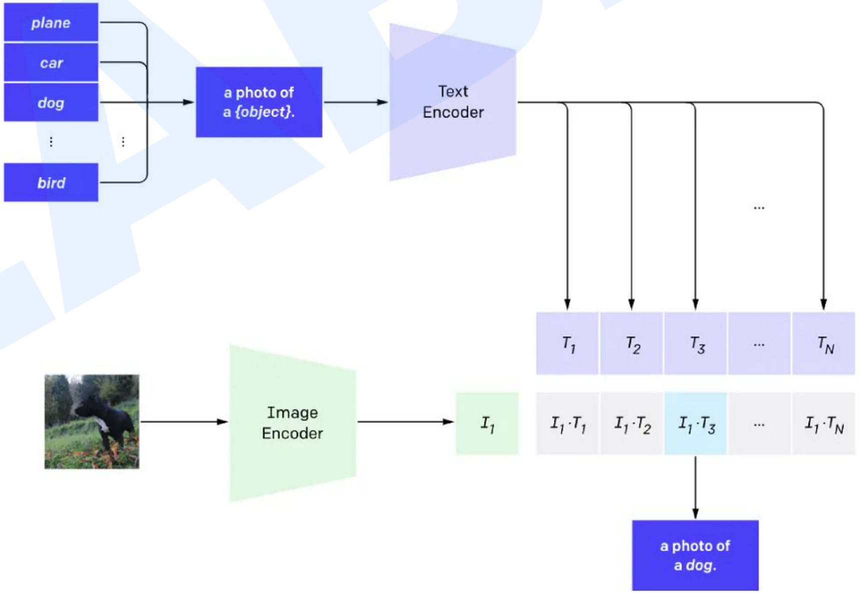
- A batch of 32,768 pairs of image and text is passed through the text and image encoders simultaneously to generate the vector representations of the text and the associated image, respectively.



User cases: Image Search With CLIP

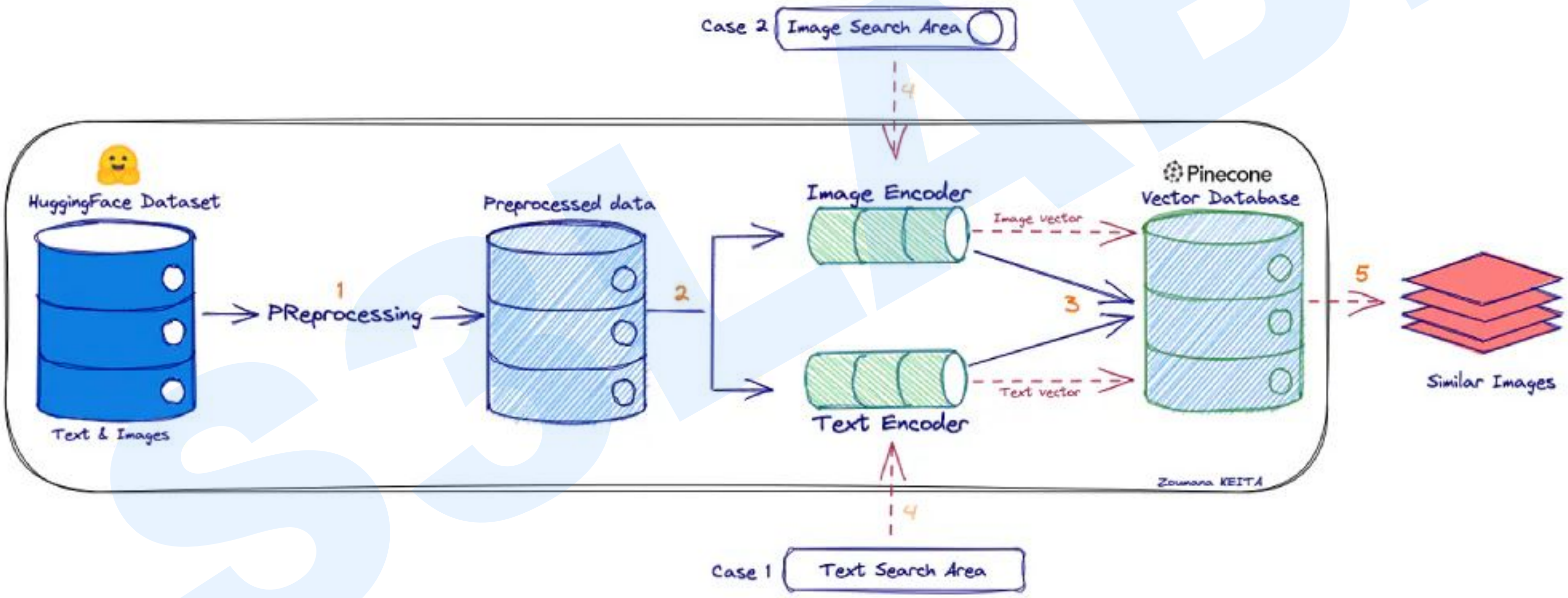
How it works - dataset classifier from label text and Zero shot detection

- Encodes all the labels/objects in the following context format: “a photo of a {object}.
- Predict which image vector corresponds to which context vector.



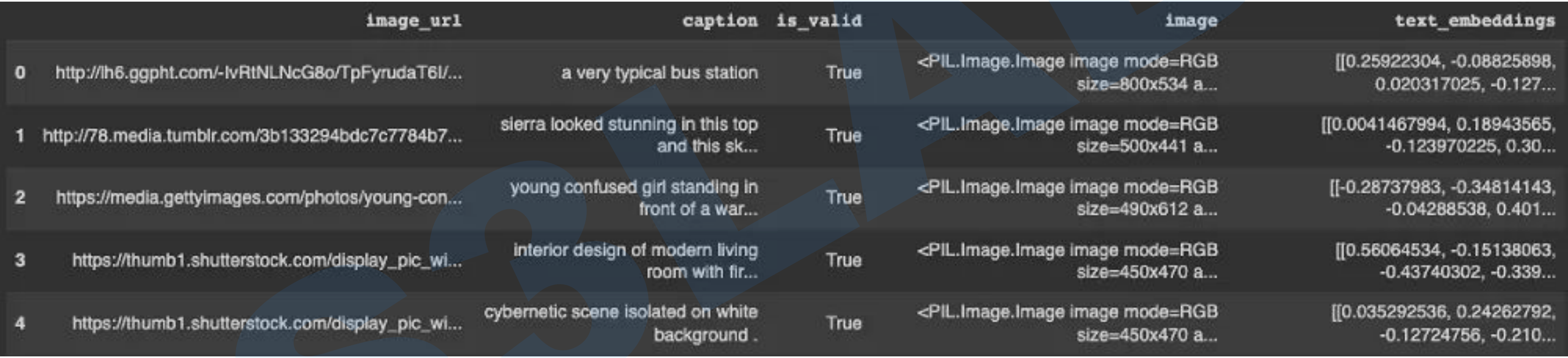
User cases: Image Search With CLIP

How it works - General workflow



User cases: Image Search With CLIP

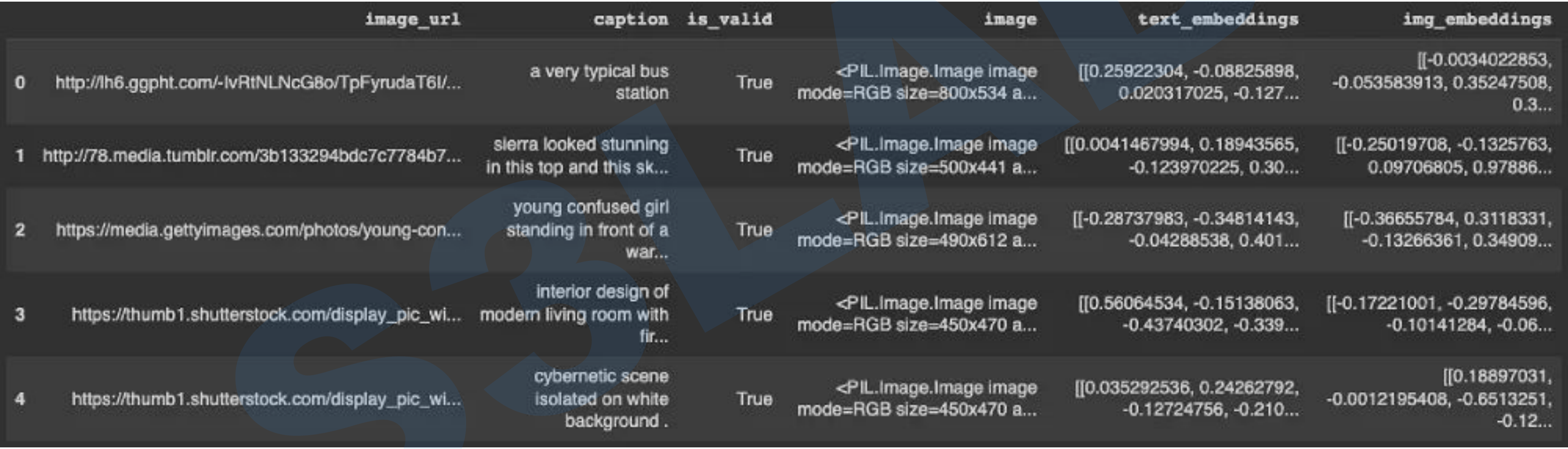
How it works - General workflow



	image_url	caption	is_valid	image	text_embeddings
0	http://lh6.ggpht.com/-IvRtNLNcG8o/TpFyrudaT6I/...	a very typical bus station	True	<PIL.Image.Image image mode=RGB size=800x534 a...	[[0.25922304, -0.08825898, 0.020317025, -0.127...
1	http://78.media.tumblr.com/3b133294bdc7c7784b7...	sierra looked stunning in this top and this sk...	True	<PIL.Image.Image image mode=RGB size=500x441 a...	[[0.0041467994, 0.18943565, -0.123970225, 0.30...
2	https://media.gettyimages.com/photos/young-con...	young confused girl standing in front of a war...	True	<PIL.Image.Image image mode=RGB size=490x612 a...	[[-0.28737983, -0.34814143, -0.04288538, 0.401...
3	https://thumb1.shutterstock.com/display_pic_wi...	interior design of modern living room with fir...	True	<PIL.Image.Image image mode=RGB size=450x470 a...	[[0.56064534, -0.15138063, -0.43740302, -0.339...
4	https://thumb1.shutterstock.com/display_pic_wi...	cybernetic scene isolated on white background .	True	<PIL.Image.Image image mode=RGB size=450x470 a...	[[0.035292536, 0.24262792, -0.12724756, -0.210...

User cases: Image Search With CLIP

How it works - General workflow



	image_url	caption	is_valid	image	text_embeddings	img_embeddings
0	http://lh6.ggpht.com/-IvRtNLNcG8o/TpFyrudaT6I/...	a very typical bus station	True	<PIL.Image.Image image mode=RGB size=800x534 a...	[[0.25922304, -0.08825898, 0.020317025, -0.127...	[[-0.0034022853, -0.053583913, 0.35247508, 0.3...
1	http://78.media.tumblr.com/3b133294bdc7c7784b7...	sierra looked stunning in this top and this sk...	True	<PIL.Image.Image image mode=RGB size=500x441 a...	[[0.0041467994, 0.18943565, -0.123970225, 0.30...	[[-0.25019708, -0.1325763, 0.09706805, 0.97886...
2	https://media.gettyimages.com/photos/young-con...	young confused girl standing in front of a war...	True	<PIL.Image.Image image mode=RGB size=490x612 a...	[[-0.28737983, -0.34814143, -0.04288538, 0.401...	[[-0.36655784, 0.3118331, -0.13266361, 0.34909...
3	https://thumb1.shutterstock.com/display_pic_wi...	interior design of modern living room with fir...	True	<PIL.Image.Image image mode=RGB size=450x470 a...	[[0.56064534, -0.15138063, -0.43740302, -0.339...	[[-0.17221001, -0.29784596, -0.10141284, -0.06...
4	https://thumb1.shutterstock.com/display_pic_wi...	cybernetic scene isolated on white background .	True	<PIL.Image.Image image mode=RGB size=450x470 a...	[[0.035292536, 0.24262792, -0.12724756, -0.210...	[[0.18897031, -0.0012195408, -0.6513251, -0.12...

References

- https://weaviate.io/blog/what-is-a-vector-database
- https://nthu-datalab.github.io/db/slides/20_Vector-DBMS.pdf
- https://www.v7labs.com/blog/vector-databases#the-future-of-vector-databases-in-ai
- https://www.pinecone.io/learn/vector-database/

Q & A

Thank you for listening

*"Coming together is a beginning;
Keeping together is progress;
Working together is success."*
- HENRY FORD