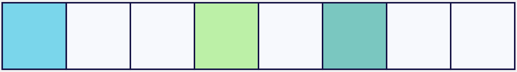
6 types of

Embeddings

and when to use them

For AI Applications

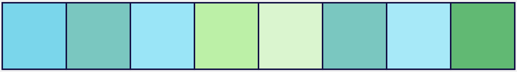
Weaviate



Sparse embeddings

Sparse vectors are often high-dimensional with **many zero values**.

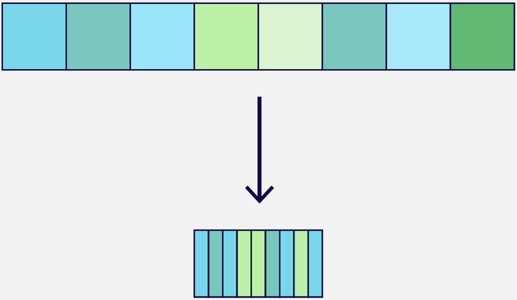
They are generated from algorithms like BM25 and SPLADE and are used in **keyword-based search**.



Dense embeddings

Dense embeddings contain mostly **non-zero values** and are generated from machine learning models like Transformers.

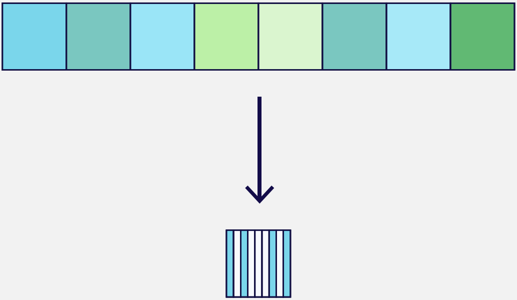
These vectors capture the semantic meaning of text and are used in **semantic search**.



Quantized embeddings

Compressed dense vectors using **lower-precision data types** (e.g., float32 to int8).

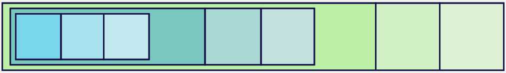
Reduces memory usage and speeds up search while maintaining most semantic information.



Binary embeddings

Extreme quantization, reducing vector components to **binary (0 or 1) values**.

Drastically reduces memory use.



Variable dimensions

Flexible embedding sizes, like **Matryoshka embeddings**.

Encode information hierarchically, allowing adaptation to different tasks or computational constraints while preserving semantic meaning.

Multi-vector embeddings

Usage of **multiple vectors instead of one pooled vector** to represent e.g., token-wise embeddings (e.g., ColBERT).

Allows for more detailed representation of complex texts.