Google Developer Group
On Campus FST

Google Developer Group
on Campus SUP'COM

# Beyond the Database : Scaling Vector Search with Qdrant

Presented to you by Mohamed Arbi Nsibi

# Mohamed Arbi Nsibi

- ML engineer
- Qdrant Star ⭐
- Former GDSC Lead 23/24

Qdrant

# What is a ⭐anyway?

Qdrant

# What is a ★anyway?

A program for developers building, sharing, and leading in the Qdrant community.

# What is  Qdrant?

Qdrant

# What Qdrant is not?

Qdrant

# Qdrant is not a Database

Andrey Vasnetsov:
Qdrant CTO

Qdrant

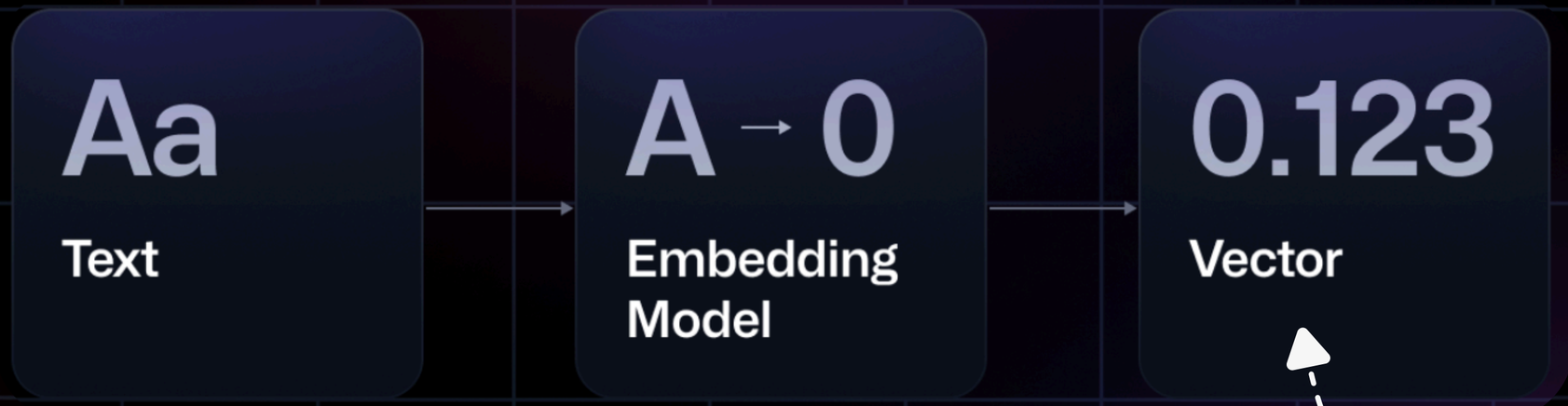
# Qdrant is not a Database

**Because Vectors are not data**

Qdrant


**Doesn’t create new information**

Qdrant




**Doesn’t create new information**

mathematical representation of meaning

Qdrant

# There is no Universal Data Storage

Qdrant

# There is no Universal Data Storage

Traditional DBs:
structured data

Unstructured data
(images and natural language)

# Qdrant is ...

## a search engine

# Qdrant is ...

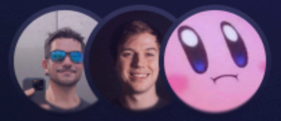
## engineered for vectors

Qdrant

# Qdrant is ...

Fully Open-source
Self-hosting : run on your own infra
Super fast: Latency  ~0.024s (~24ms)
Hybrid Search
UI support
Free Tier ~1M(vectors) 768-dim

60K



Community Members

250M+

OSS Downloads

26K+

Github Stars

>140

Contributors

Qdrant

# The Shift to AI Native Search

Qdrant

Unstructured Data Is Exploding

(Data isn't in a spreadsheet)

AI Agents Are the New Users

Legacy Search Falls Short

Vector Search Is the Missing Layer

## Wave 1

RAG 1.0 - Static Assistants

(2023 - 2024)

## Wave 2

Agentic AI - Multi-Step Reasoning

(2024 - Now)

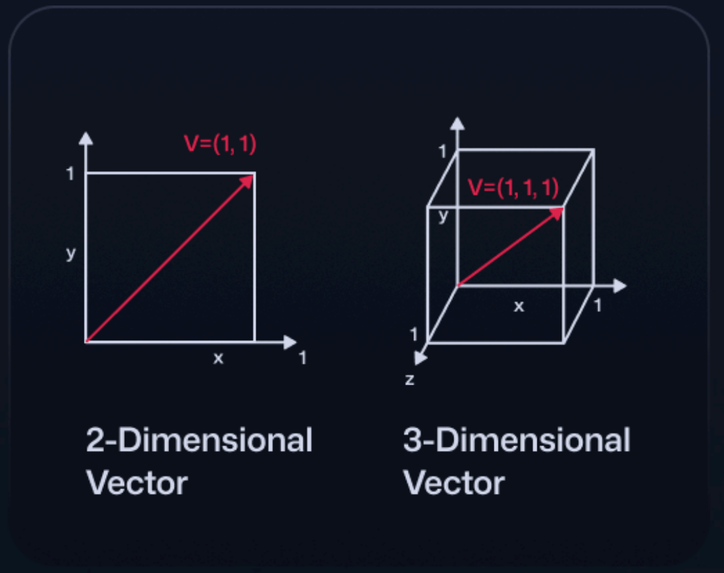
## Wave 3

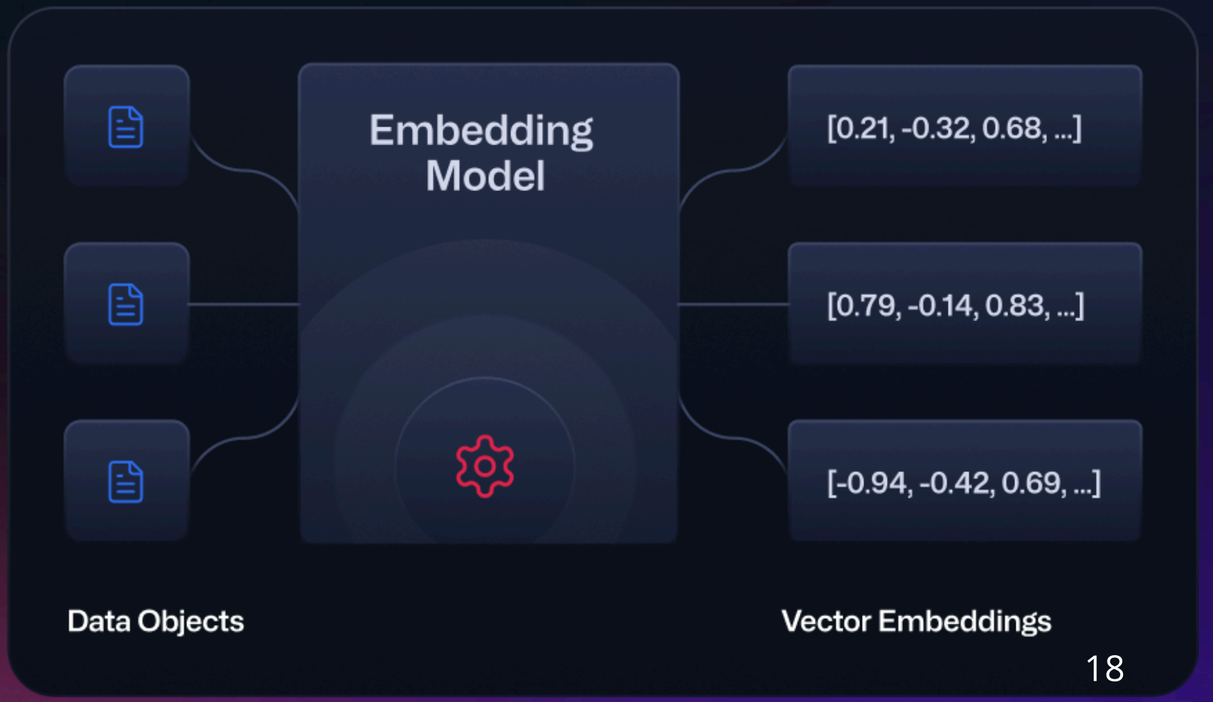
Embedded AI – Physical & On-Edge Agents

(2025 + )

# Vector Search Basics

Two different vector embeddings should be close to each other if they represent a similar input object.
Embeddings are generated by neural networks and can represent thousands of dimensions.



# Vector Search Basics

Although word counting produces embeddings, dense embeddings are needed to capture semantics

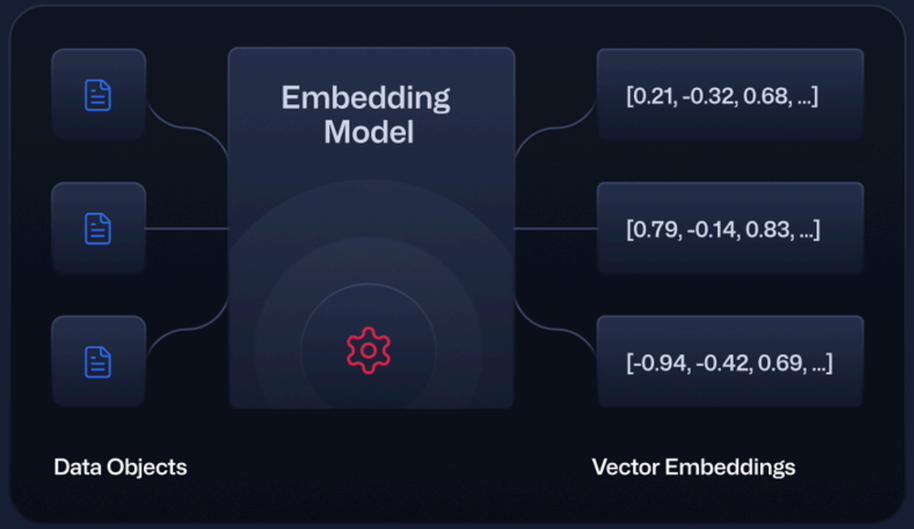
## Sparse embedding:
*e.g. One Hot Encoding*

| | an | another | embedding | is | this | Query Sim. |
|---|---|---|---|---|---|---|
| *“this is an embedding”* | [1, | 0, | 1, | 1, | 1] | 3 |
| *“this is another embedding”* | [0, | 1, | 1, | 1, | 1] | 2 |

***Query:***

***“What is an embedding?”***

## Dense embedding:
*e.g. from BERT*




Gemini TwelveLabs NOMIC

# Qdrant-at-a-Glance

Vector Search Engine. Not Database. optimized for scalability and high availability

### Built-Out for Search-First Workflows

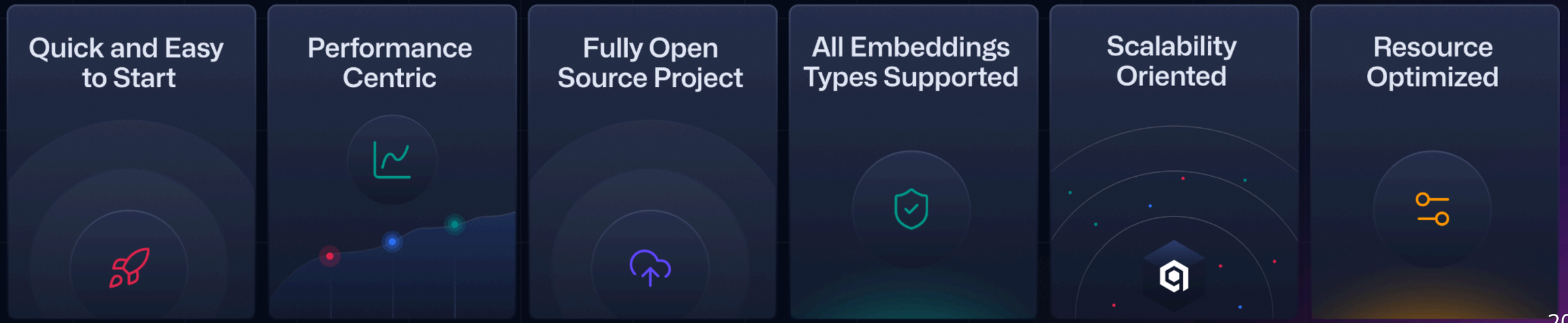
Qdrant is built from the ground up with **search as the core functionality**. Conventional databases focus on ACID transactions and strong consistency.
In contrast, search engines are optimized for scalability, low-latency search, and high availability.

### Engineered for Vector Search at Scale

Qdrant is purposed to handle extremely high-dimensional embeddings. It's designed with a **vector index as a central component of the system**, allowing a custom, finely tuned approach to data and index management that secures high performance even as data grows and changes dynamically

### Specialized for Advanced Vector Operations

Qdrant is designed from the ground up to handle high-dimensional vector math and (dis-)similarity-based retrieval. This allows for leveraging the full potential of vector search **beyond simple similarity ranking** from multi-stage filtering to dynamic exploration of high-dimensional spaces**.**



# How Qdrant Achieves Search

## Core Capabilities

### Vector Search
Scalable similarity and discovery search (billions of vectors)

### Hybrid Search
Combine dense + sparse embeddings, filters, and metadata

### Filtering
Numeric, categorical, geo, temporal filters out-of-the-box

### Distributed & Resilient
Replication, sharding, multi-tenancy

## Advanced Features

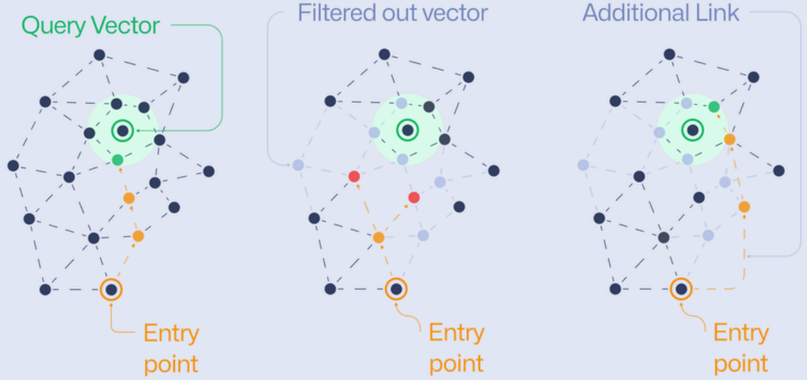
### Re-ranking
Maximum Marginal Relevance (MMR), score boosting

### Quantization
Binary, scalar & product; lower cost without major recall loss

### Multi-vectors: Late interaction for retrieval models (e.g. ColBERT)

### Performance Optimizations
HNSW tuning, payload indexing, prefetching

Filterable HNSW




Similarity Search

Similarity Search with MMR

# Qdrant Innovations to make development easier

## FastEmbed

Generate high-quality embeddings fast. A small Python library for embedding generation, built in and integrated with Qdrant.

- Works out of the box in Qdrant.
- Few dependencies: runs on CPU; skips multi-GB PyTorch downloads.
- Made for speed: uses ONNX Runtime and data parallelism.

**Key features**

- Use Qdrant models (miniCOIL, BM42).
- Support for late-interaction (ColPali, ColBERT) and sparse-neural methods (SPLADE, BM42, miniCOIL), MUVERA embeddings and more.
- Run inference and upsert/search in one call.

**Import:**

```
from qdrant_client.models import
Document, Image
```

## MCP Servers

Build custom retrieval-based AI apps fast. Start from these servers and add tools/commands for your data and workflows.

- mcp-server-qdrant: official MCP server for storing and retrieving data in Qdrant.
- mcp-for-docs: open-source API reference for AI coding assistants using semantic code retrieval.

**Key features**

- Automate codebase documentation.
- Personalize your coding assistant to your project's conventions and rules.
- Do `inline RAG`.
- Speaks stdio, sse, and streamable-http protocols.

**Run:**

```
docker run  mcp-server-qdrant
```

## Qdrant Edge

Bring vector search to the edge: an embeddable, high-performance engine that runs directly on mobile and other edge devices.

- Run on low-CPU devices.
- Use one API to manage and synchronize data on-device and in your cloud cluster.
- Fit common on-device cases: phones and laptops, smart-home/IoT, robotics.

**Key features**

- Use local storage to avoid network latency.
- Support multi-tenant setups; treat each device as its own tenant.
- Embed as a library; runs in-process with no background daemons.

**Use:**

```
client =
QdrantClient(path="qdrant_edge.db")
```

# Vector Search in Production

- Written in **Rust** and offers **great performance**
- Allows to interact by **HTTP** or **gRPC protocols**.
- Runs both in **single and multiple node** setup.
- Incorporates **category, geo-coordinates** and **full-text filters**
- Supports **hybrid**, **multimodal**, **multivector** and **multi-staged** search
- Official **Python**, **Javascript/Typescript**, **Rust** and **Go SDKs**.
- Makes vector search **affordable**.

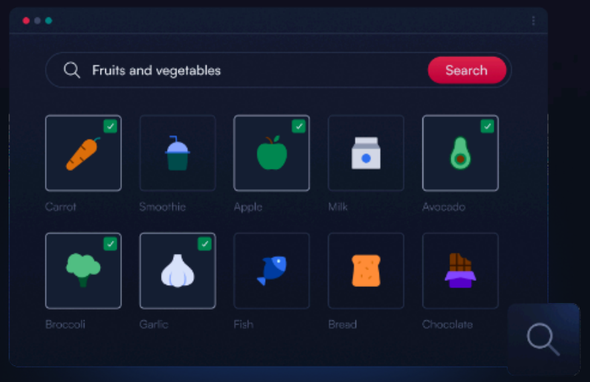
For Managed Cloud solutions, check out Cloud Embeddings Inference.

# Vector Search

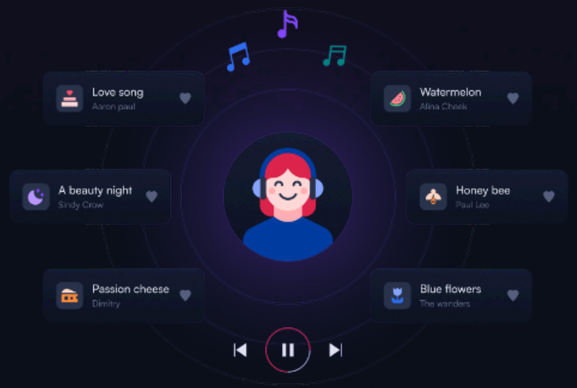
An essential part of the AI Transformation

HR-Tech | Ad-Tech | Online Dating | Gig Economy | E-Commerce | Legal-Tech | Fashion | Med-Tech

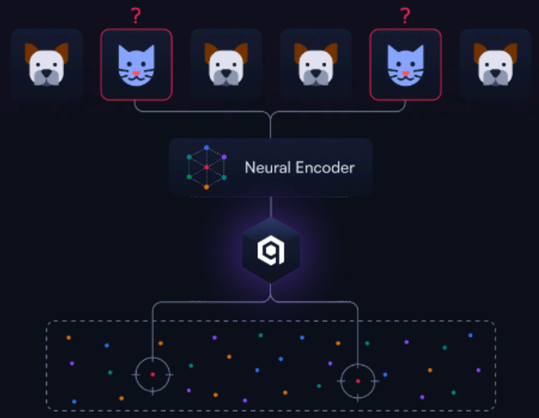
Ed-Tech | Media & News | Agriculture | Manufacturing | Streaming Services | Marketplaces | Anti-fraud | Biometrics




**Search Systems**




**Recommendations**




**Anomaly Detection**


**RAG / Information Assistants**

# Getting Started with Qdrant

### Qdrant Open Source
Usually deployed with Docker containers. Lightweight, offers all the functionalities of Qdrant.

### Qdrant Managed Cloud
Run on one of the three major cloud providers: AWS, Azure, or GCP. Provides a management UI and API. For US regions, we offer Cloud Inference that processes raw data into vectors.

### Qdrant Hybrid Cloud
All the benefits of cloud deployment, but keeping the data on your premises. Requires a Kubernetes cluster and might be managed from Qdrant Cloud UI, but no data leaves your environment.

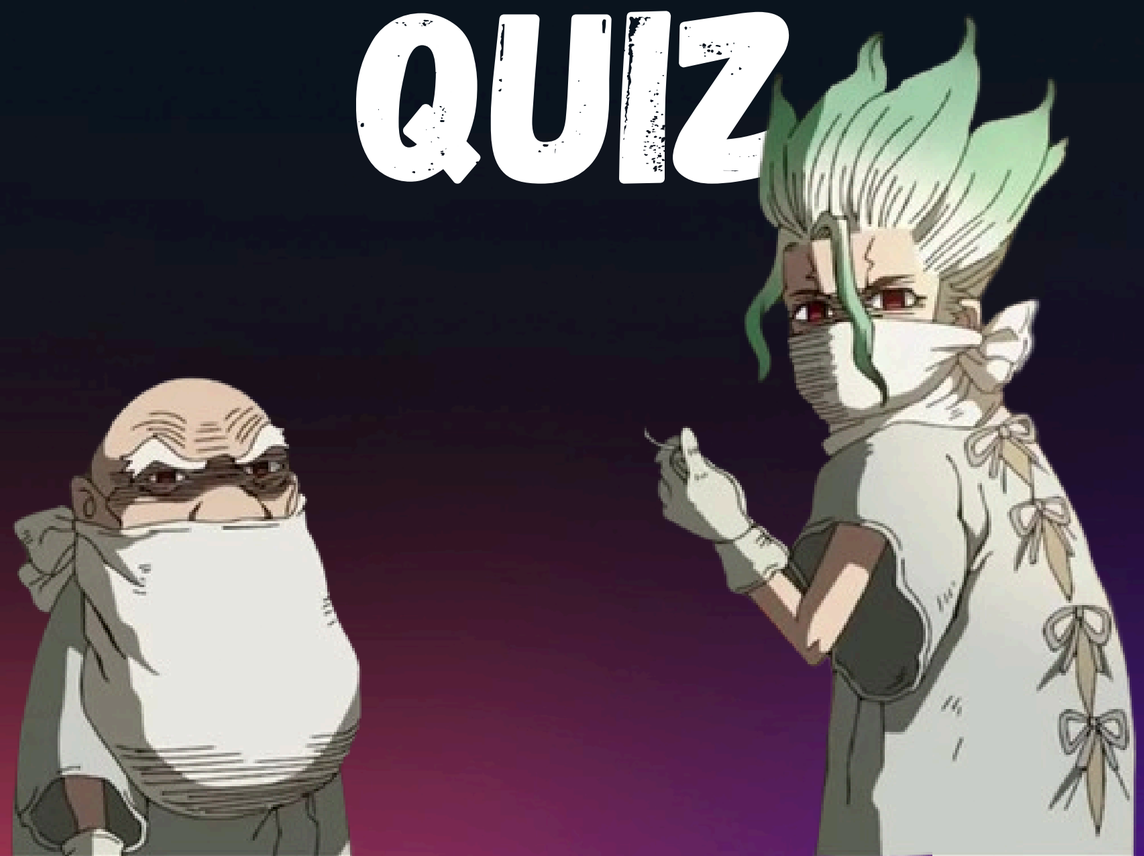
### Qdrant Private Cloud
A dedicated, on-premise solution that guarantees supreme data privacy and sovereignty.

### Python SDK Local Mode
Suitable mostly for quick experiments, but not intended to be running in production.

# Ecosystem

Gemini | n8n | MISTRAL AI_
aws | UNSTRUCTURED | A
Google Cloud Platform | OpenAI | jina
Haystack by deepset | crewai | LlamaIndex
LangChain | cohere | *and more...*

# QUIZ



# THANK YOU FOR YOUR ATTENTION!!


## Where to find me?

Goodnight

Linkedin Profile

MedArbiNsibi