Google Developer Group
On Campus FST

Google Developer Group
on Campus SUP'COM

Navigating the Qdrant Ecosystem from Local to Cloud

Presented to you by Mohamed Arbi Nsibi

Mohamed Arbi Nsibi

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

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

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.

Creating a collection/table:

```
PUT /collections/rentals
{
  "vectors": {
    "size": 300,
    "distance": "Cosine"
  }
}
```

SQL equivalent needs a schema

```
CREATE TABLE rentals (
id INTEGER PRIMARY KEY, vector FLOAT[], city TEXT,
sqft INTEGER, img_url TEXT, tags TEXT[], description TEXT
);
```

Field indexing:

```
// Vector indexing happens by default
// Each payload index adds more links to keep the graph connected for effective filtering
// Repeat for 'sqft' field with 'integer' type
PUT /collections/rentals/index
{
    "field_name": "city",
    "field_schema": {
        "type": "keyword",
    }
}
```

```
CREATE INDEX rentals_city_sqft_idx ON rentals (city, sqft);
```

Search/Read:

```
POST /collections/rentals/points/search
{
  "query": [0.2, 0.3, ..., 0.4], // generated from user query (text) using same model
  "filter": { "must": [ {"key": "city", "match": {"value": "Bangalore"}}, {"key": "sqft", "range": { "gte": 1000 }}]},
  "limit": 10
}
// Response:
[
{"id":  4,  "score":  0.56,  "payload":  {...}},
{"id":  2,  "score":  0.40,  "payload":  {...}},
{"id": 5, "score": 0.23, "payload": {...}}, ]
```

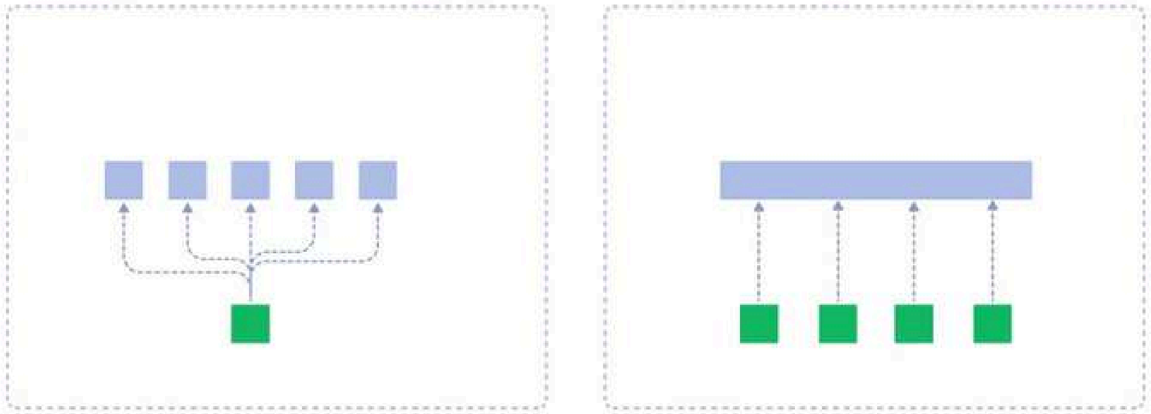
Postgres equivalent (just filtering, no vector search):

```
SELECT * FROM rentals WHERE city = 'Bangalore' AND sqft > 1000 LIMIT 10
```

Latency and Throughput

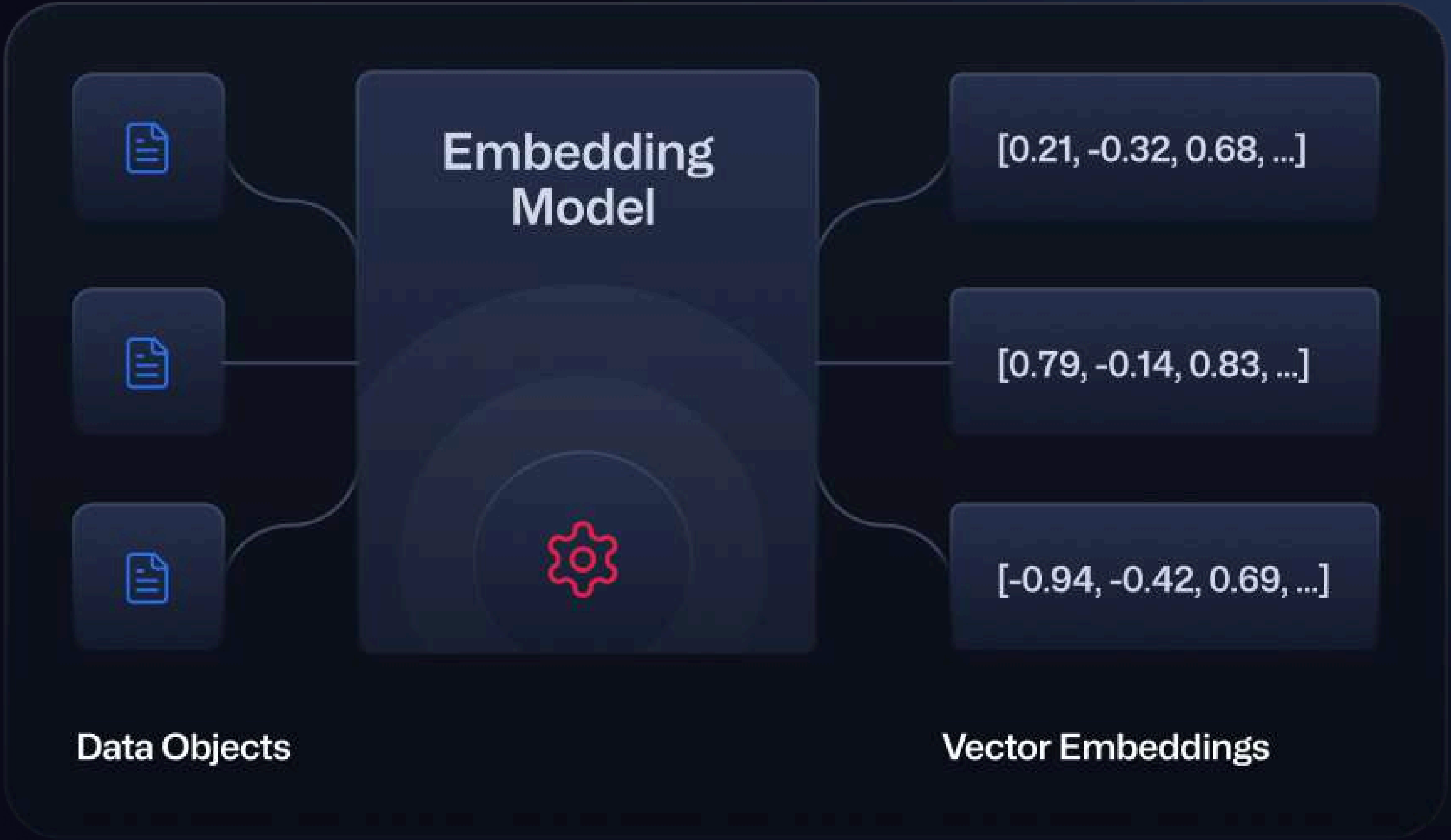
- Latency: Time taken for a single request
- Throughput: Number of requests handled / second
- Min. latency via `num_segments == num_cpu`
- Max. RPS via fewer but larger segments

But longer indexing time



Vector Search Basics

Dense embedding:
e.g. from BERT



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

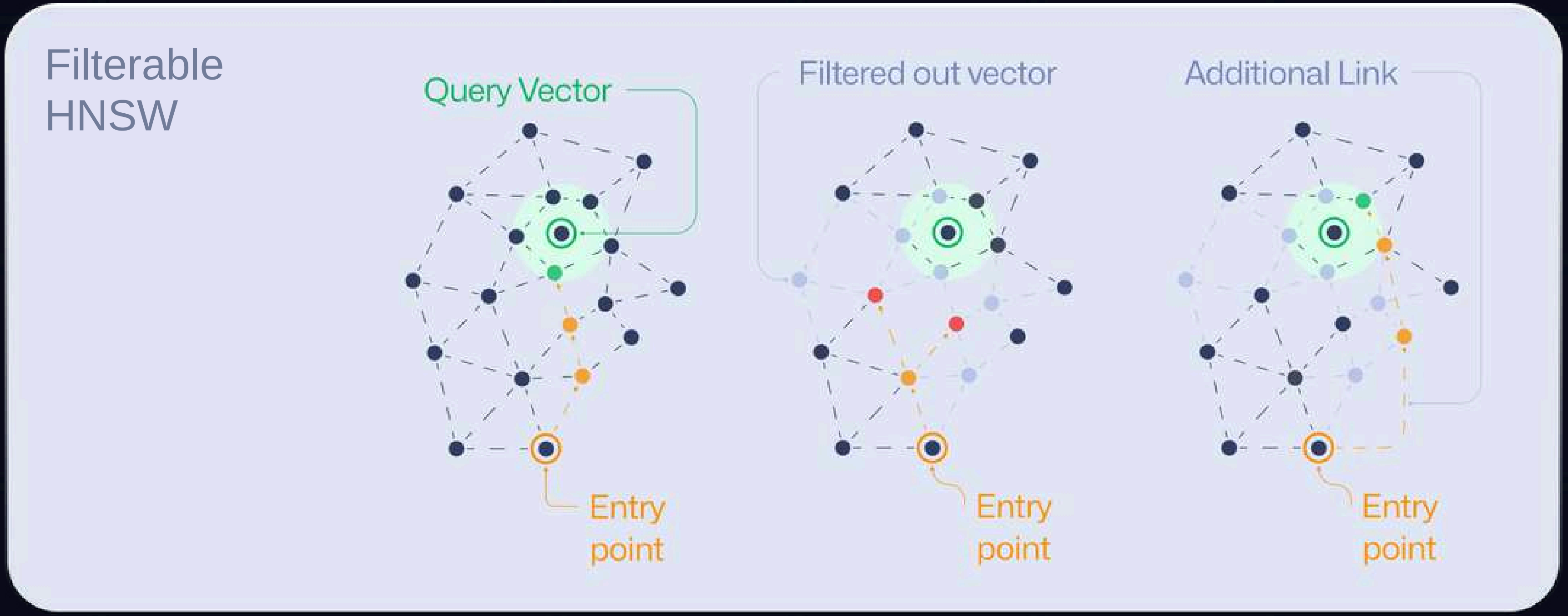
Advanced Features

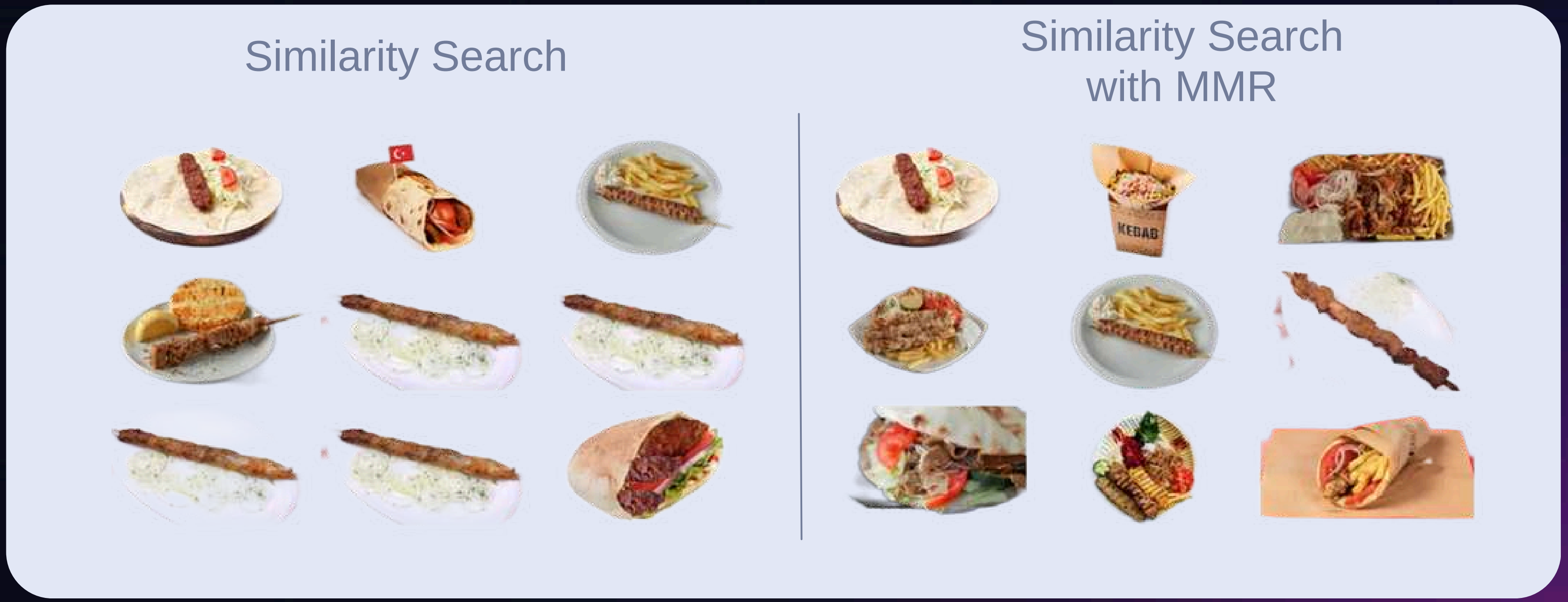
Re-ranking
Maximum Marginal Relevance (MMR), score boosting

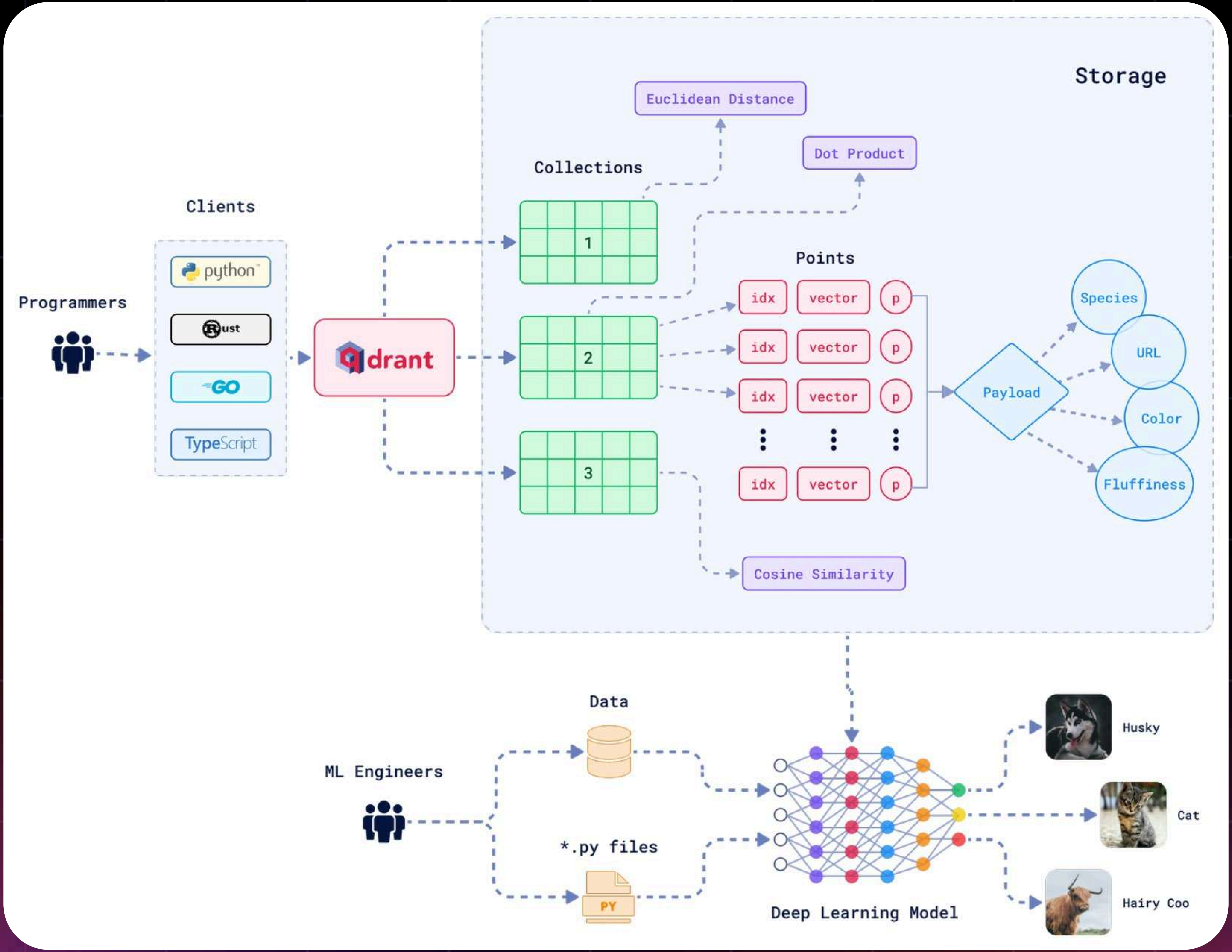
Quantization
Binary, scalar & product; lower cost without major recall loss

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

Performance Optimizations
HNSW tuning, payload indexing, prefetching






Storage
Euclidean Distance
Dot Product
Collections
Clients
Points
Programmers
python
Rust
GO
TypeScript
Qdrant
1
2
3
idx
vector
p
Payload
Species
URL
Color
Fluffiness
Cosine Similarity
Data
ML Engineers
*.py files
PY
Deep Learning Model
Husky
Cat
Hairy Coo

DEMO

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:

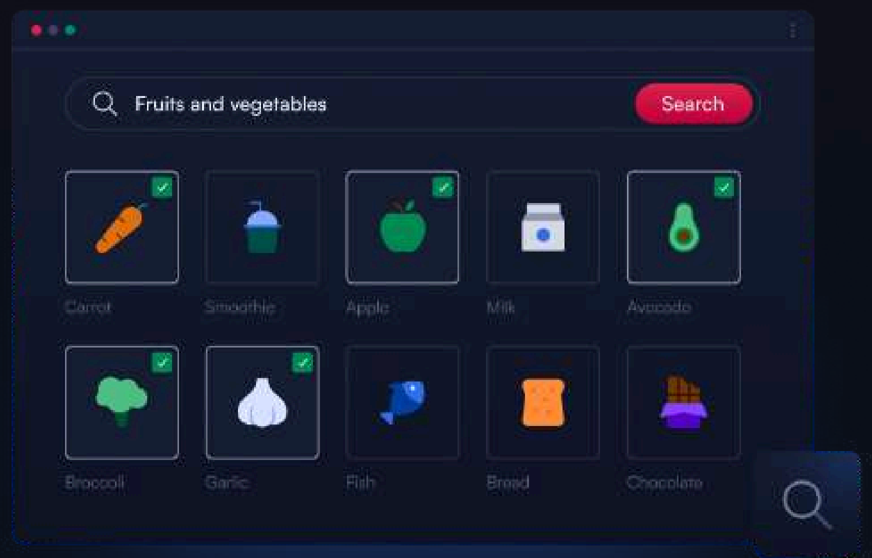
```
client =
QdrantClient(path="qdrant_edge.db")
```

Vector Search

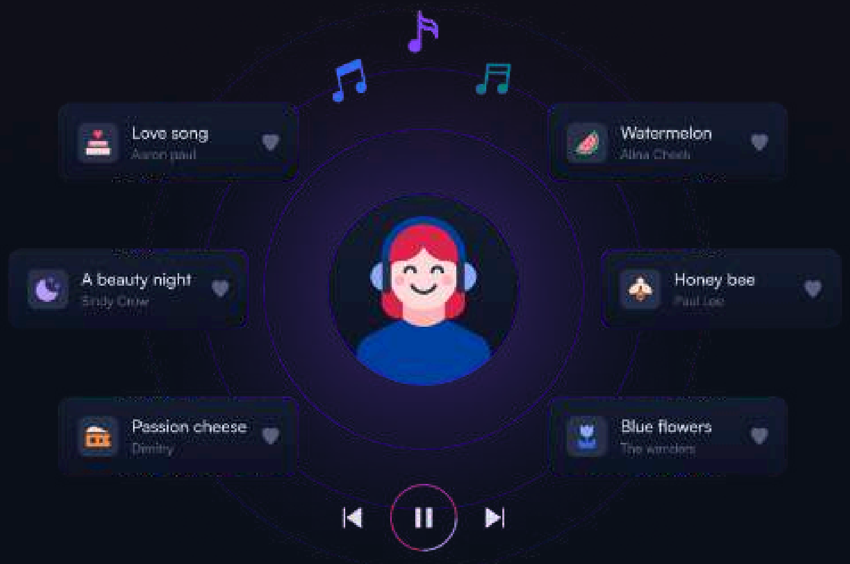
An essential part of the AI Transformation

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

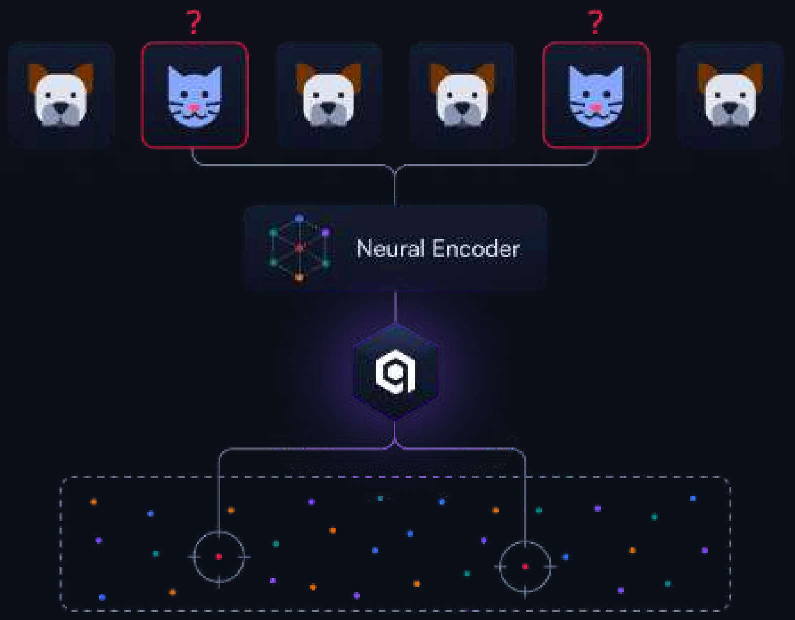
Ed-Tech | Media & News | Agriculture | Manufacturing | Streaming Services | Marketplaces | Anti-fraud | Biometrics



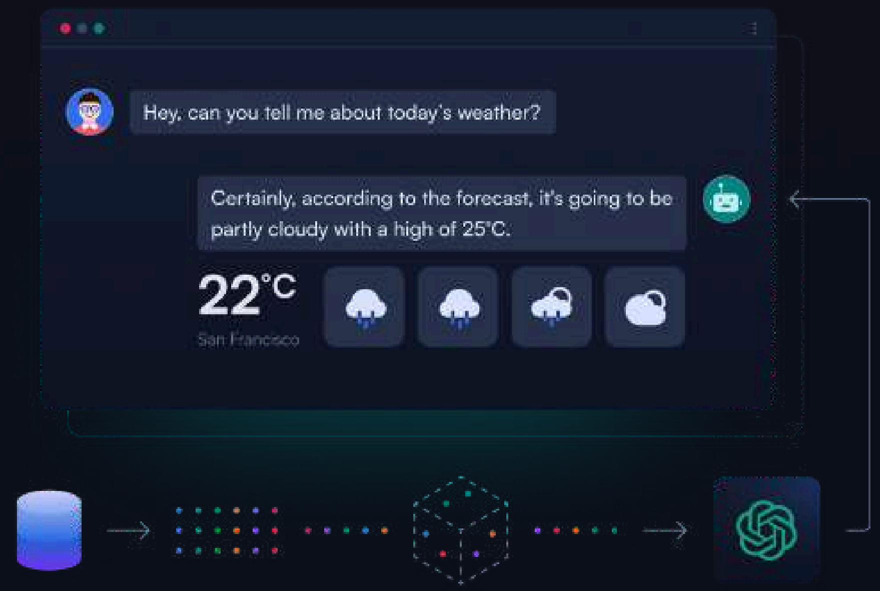
Search Systems



Recommendations



Anomaly Detection


Hey, can you tell me about today's weather?
Certainly, according to the forecast, it's going to be partly cloudy with a high of 25°C.
22°C
San Francisco

RAG / Information Assistants

LLMs providers you can start with :

Gemini

deepseek

OpenAI

Hugging Face

cohere

MISTRAL AI_

Getting Started with Qdrant

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

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.

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

Ecosystem

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

Trends in DB industry

Vector Search is a very hot paradigm right now.

- Serverless: User shouldn't know/care about the number of machines required. Generally paired with multi-tenancy.
- Neon got acquired.
- Decoupling storage and compute: Put stuff in S3, load only whatever is required.
- Often serverless. Turbopuffer, Tantivy.

Jobs in DB industry

Most DBs are open source. Pick a niche and start contributing

Internships are easier. Best way to enter via GSoC/LFX as a student for good mentorship.

- Some hot DB startups:

Vector DBs: Qdrant, LanceDB, Turbopuffer, Weaviate ,e6data,

Couchbase, Parsable, Yugabyte, Databricks, PureStorage,

Tigerbeetle, Turso, Grafana, Questdb, Supabase

60K

Community Members

250M+

OSS Downloads

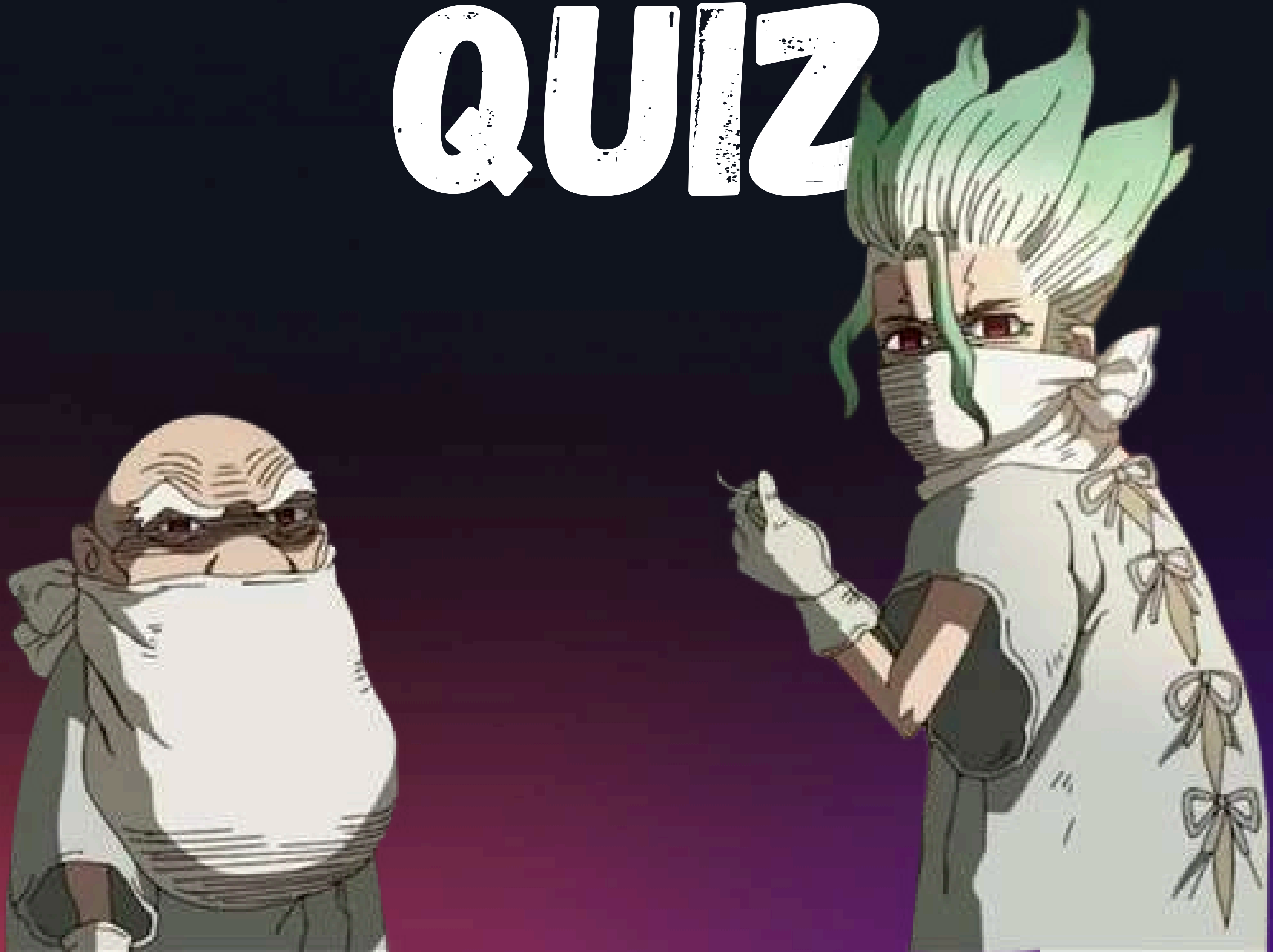
26K+

Github Stars

>140

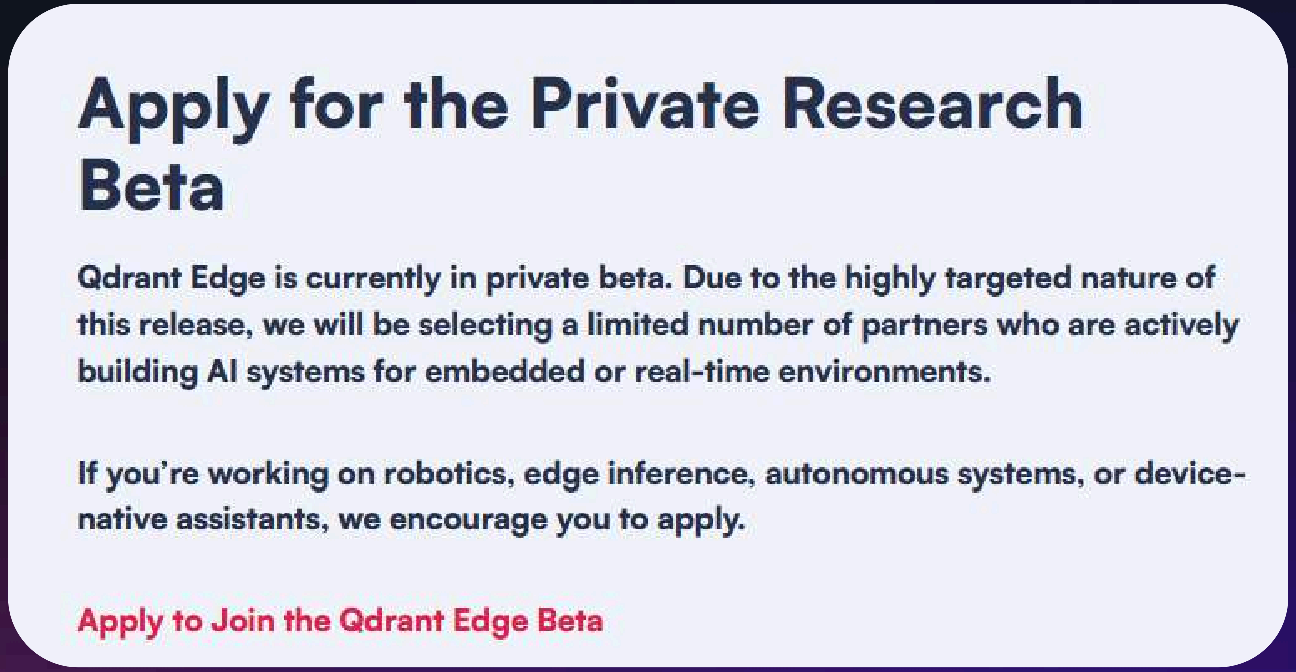
Contributors

QUIZ



Resources

- Qdrant Newsletter subscription
- Qdrant Homepage
- Qdrant Documentations
- Qdrant Cloud signup
- Just-RAG github repo
- Qdrant-resources Github Repo
- Previous Talk materials
- Qdrant On edge
- Metric learning for anomaly detection
- Music recommendation system
- Filtering with qdrant



Apply for the Private Research Beta

Qdrant Edge is currently in private beta. Due to the highly targeted nature of this release, we will be selecting a limited number of partners who are actively building AI systems for embedded or real-time environments.

If you're working on robotics, edge inference, autonomous systems, or device-native assistants, we encourage you to apply.

Apply to Join the Qdrant Edge Beta

THANK YOU FOR YOUR ATTENTION!!

Where to find me?

Goodnight

Linkedin Profile

MedArbiNsibi