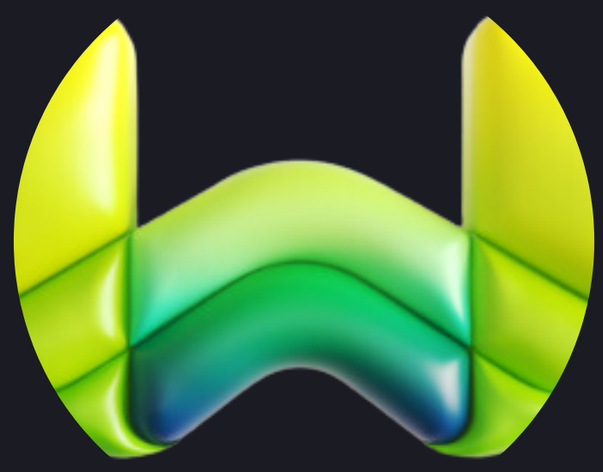
Qendel AI

# Weaviate's Guide to VECTOR QUANTIZATION

---

## Every Data Scientist Should Know



Qendel AI

Project scaling.

Data accumulating.

Momory usage growing.

Retrieval slowing.

Cost Skyrocketing.

Sounds familiar?

Qendel AI

We are in the same boat.

But, there is a solution

# VECTOR QUANTIZATION

Qendel AI

# VECTOR QUANTIZATION

✅ Cuts down memory needs

✅ Reduces latency

✅ Slashes cost

Qendel AI

# WEAVIATE

Offers solution

in 2 ways

Qendel AI

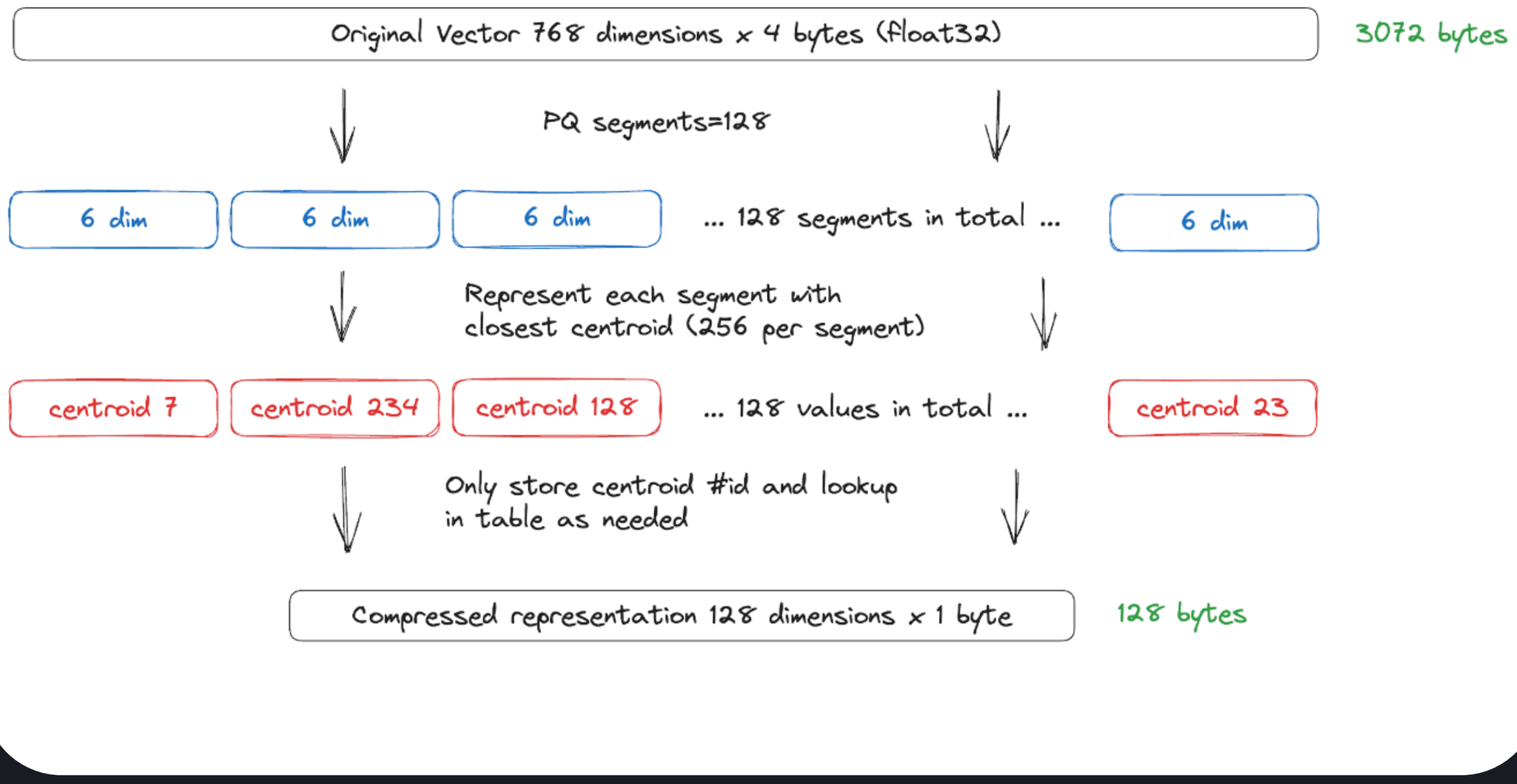
# 1

# Product Quantization (PQ)

Qendel AI

# What PQ does

Compresses your vector embeddings by breaking them down into smaller, manageable segments.



Qendel AI

# PQ Benefits

Reduces memory usage by almost 24 times while maintaining a balance between performance and recall.

---

# Best for

Those who use hnsw indexes and need a fine balance between speed and accuracy.

Qendel AI

# 2

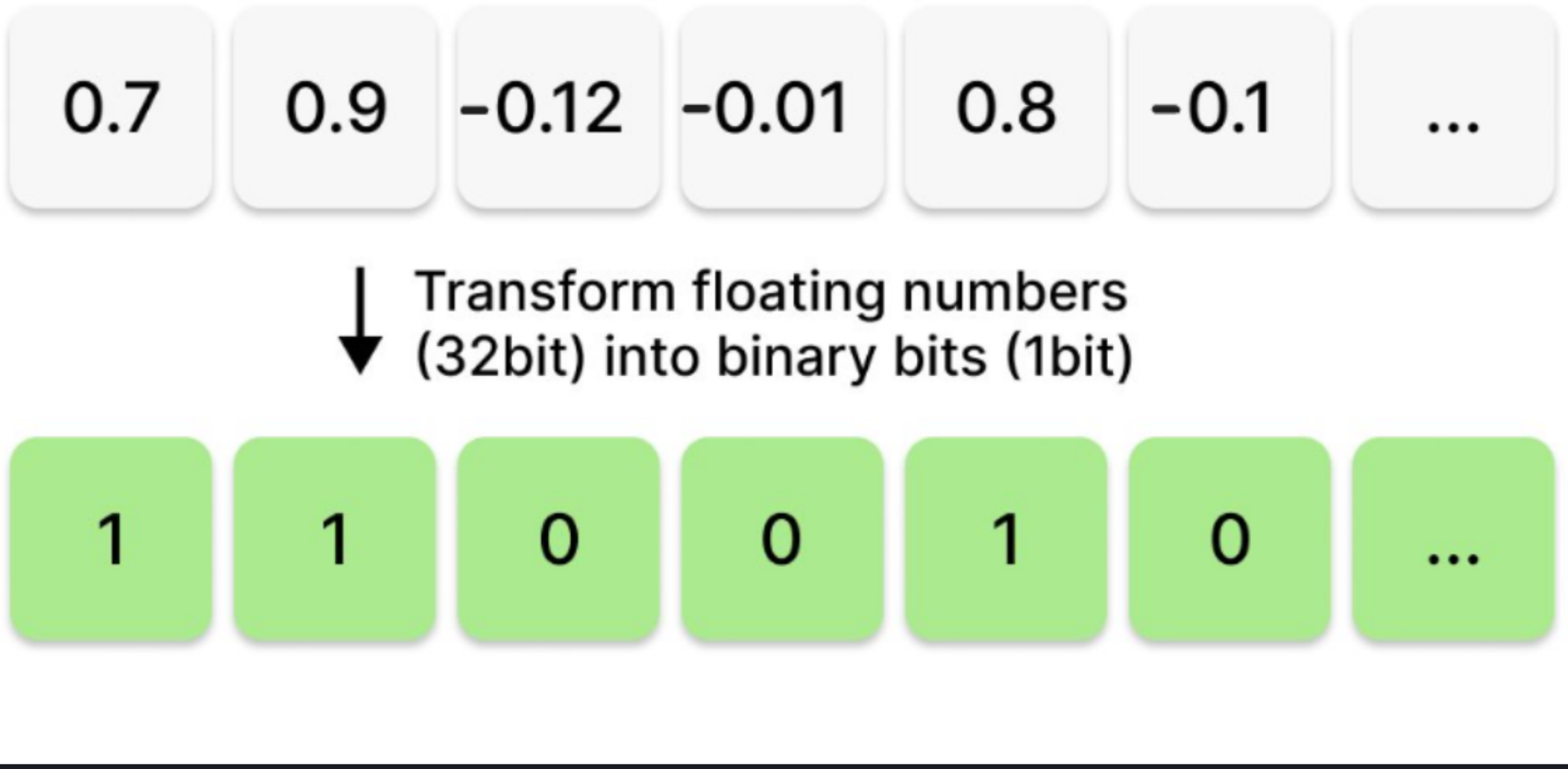
# Binary Quantization

# (BQ)

Qendel AI

# What BQ does

Converts each vector into a binary format, drastically reducing the size from bytes to bits.



# BQ Benefits

Achieves a 32x reduction in storage requirements and speeds up search processes.

---

# Best for

Projects where speed is critical, and slight compromises on accuracy are acceptable.

# 📚 Trade-offs

- **PQ** might slightly reduce recall but saves more memory.

- **BQ** offers incredible speed at the cost of some accuracy.

Qendel AI

# 📚 Bonus

## 🔗 Check links in the comment

Qendel AI

# Want more content like this?

Follow Qendel AI for daily tips on

📌 Prompting

📌 LLMs

📌 RAG

📌 Agents