



# Vector Databases and Neural Search

Dmitry Kan, Max Irwin  
Open NLP Meetup

# About me

- PhD in NLP
- 16+ years of experience in developing search engines for start-ups and multinational technology giants
- Expert in vector search engines and the host of the Vector Podcast
- Blogging about vector search on Medium:  
<https://dmitry-kan.medium.com/>
- Committer on search QA tool Quepid:  
<https://github.com/o19s/quepid>



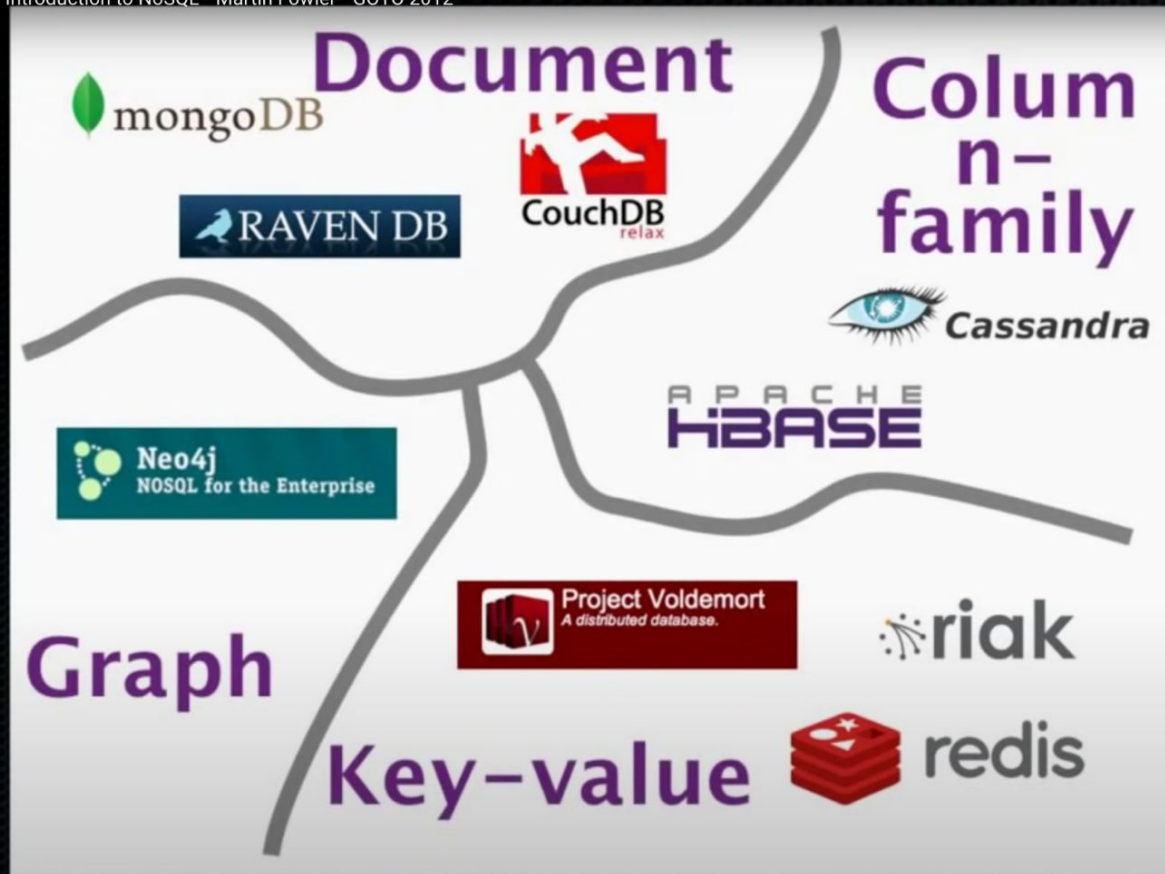
# Team Sisu: BigANN Competition @ NeurIPS'21

- Max Irwin
- Alex Semenov
- Aarne Talman
- Leo Joffe
- Alex Klibisz
- Dmitry Kan

[Billion-Scale ANN Algorithm Challenge](#)

# Scope

- Vector search in a nutshell
- Main Vector DBs
- Neural search frameworks
- Get practical
- Demos



# Google Trends

● **neural search**  
Search term

● **inverted search**  
Search term

+ Add comparison

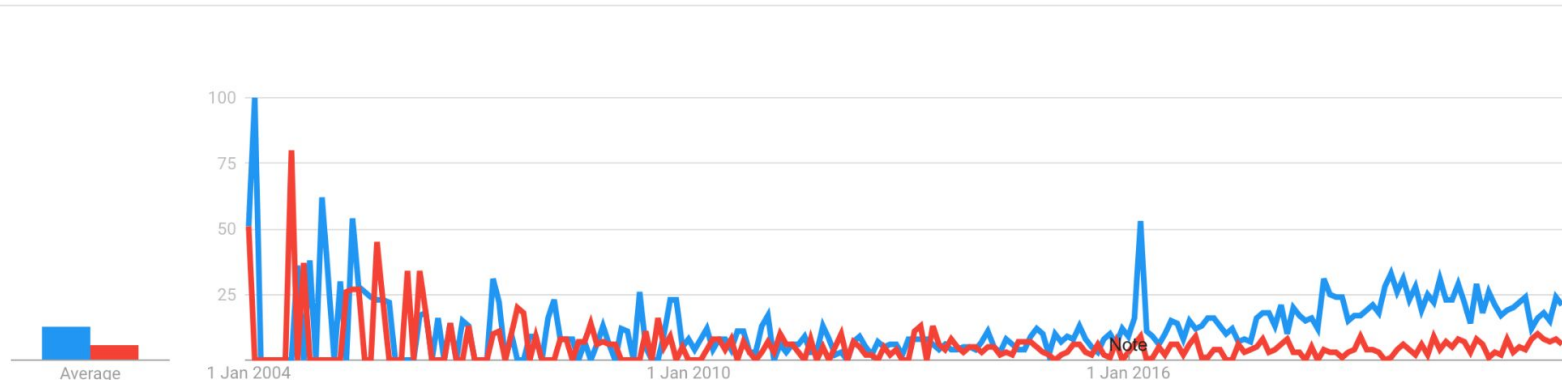
Worldwide ▼

2004 – present ▼

All categories ▼

Web Search ▼

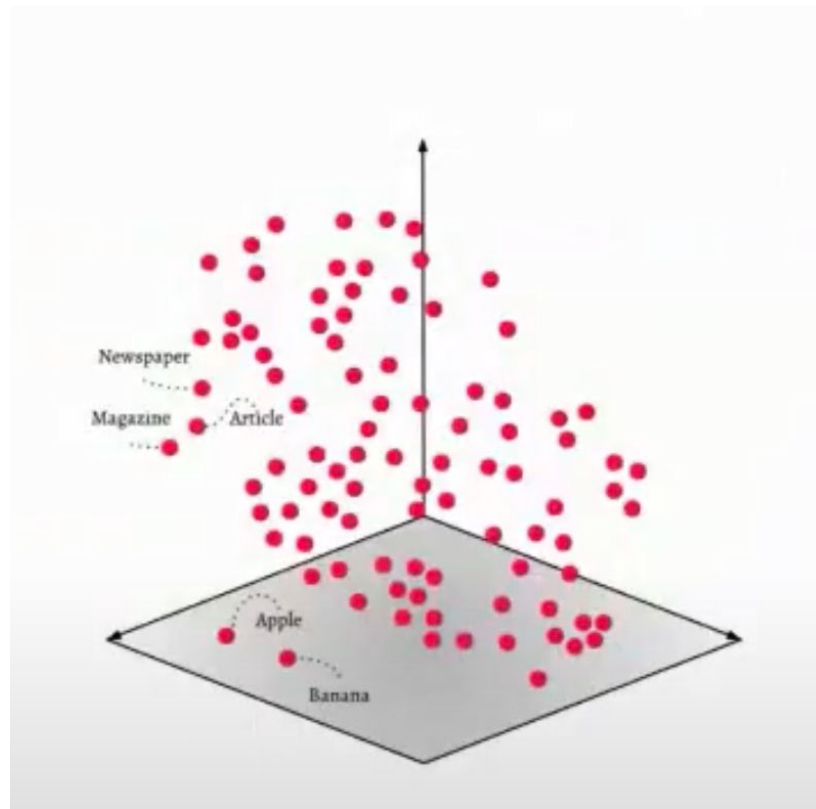
Interest over time ?



# Vector search in a nutshell

**Vector search** is a way to represent and search your objects (documents, songs, images..) in a geometric space (usually of high-dimension) in the form of an embedding (a vector of numbers:  $[0.9, -0.1, 0.15, \dots]$ )

- At small scale you can apply exact KNN search
- At larger scale you need to use ANN search: trade some precision for speed



Credit: Weaviate V1.0 release - virtual meetup

## Use cases

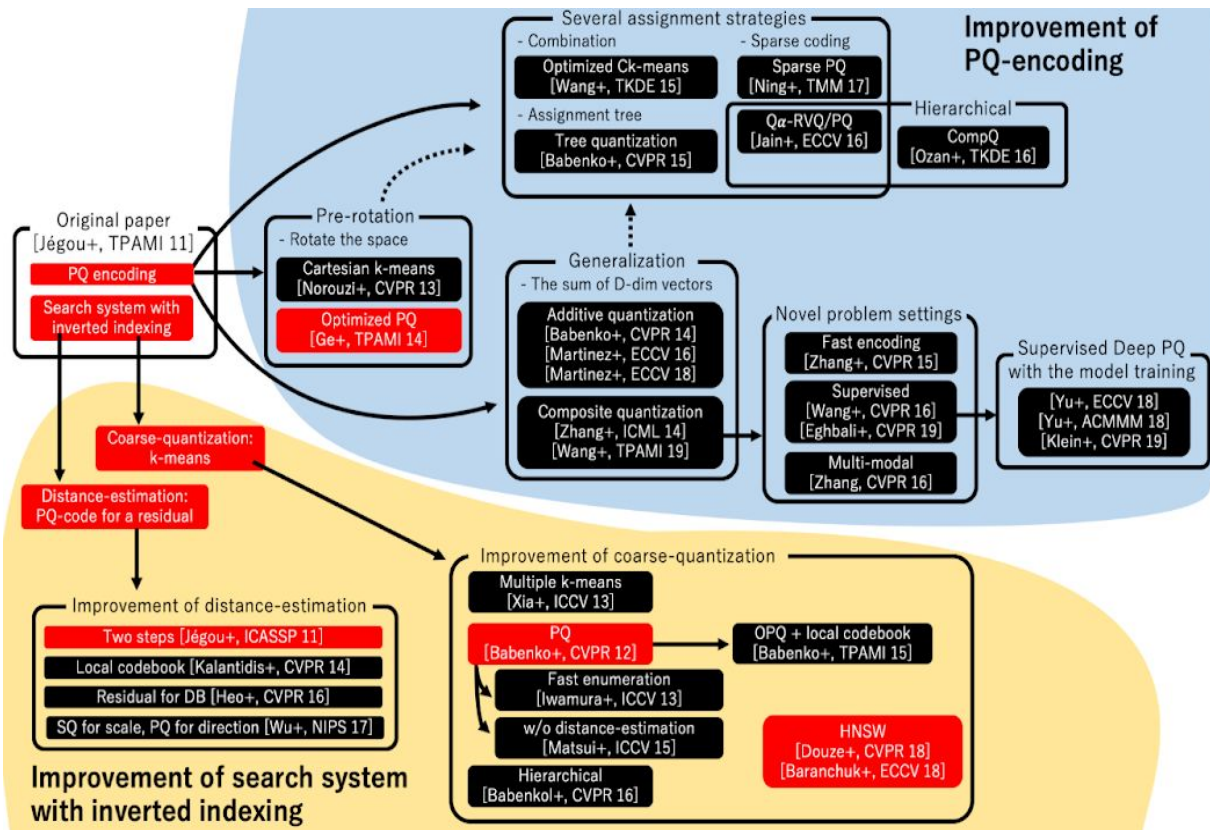
- Semantic search
- Image similarity
- Sound search
- Multimodality: searching images with text
- Recommenders
- E-commerce zero-hit long-tail (similarity search)



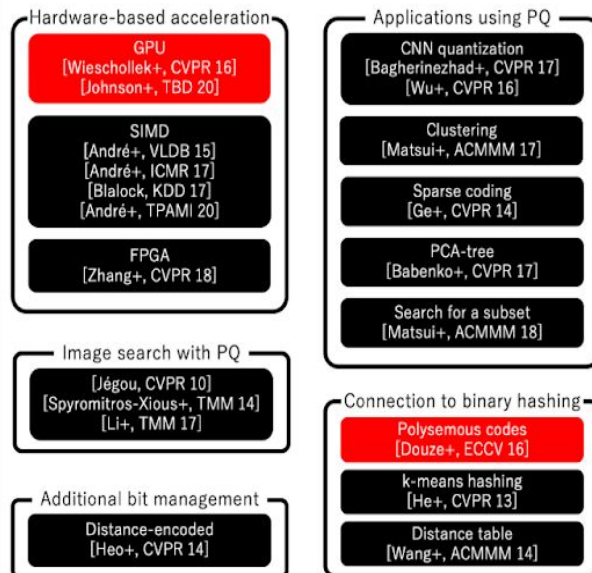
# Big players in the game

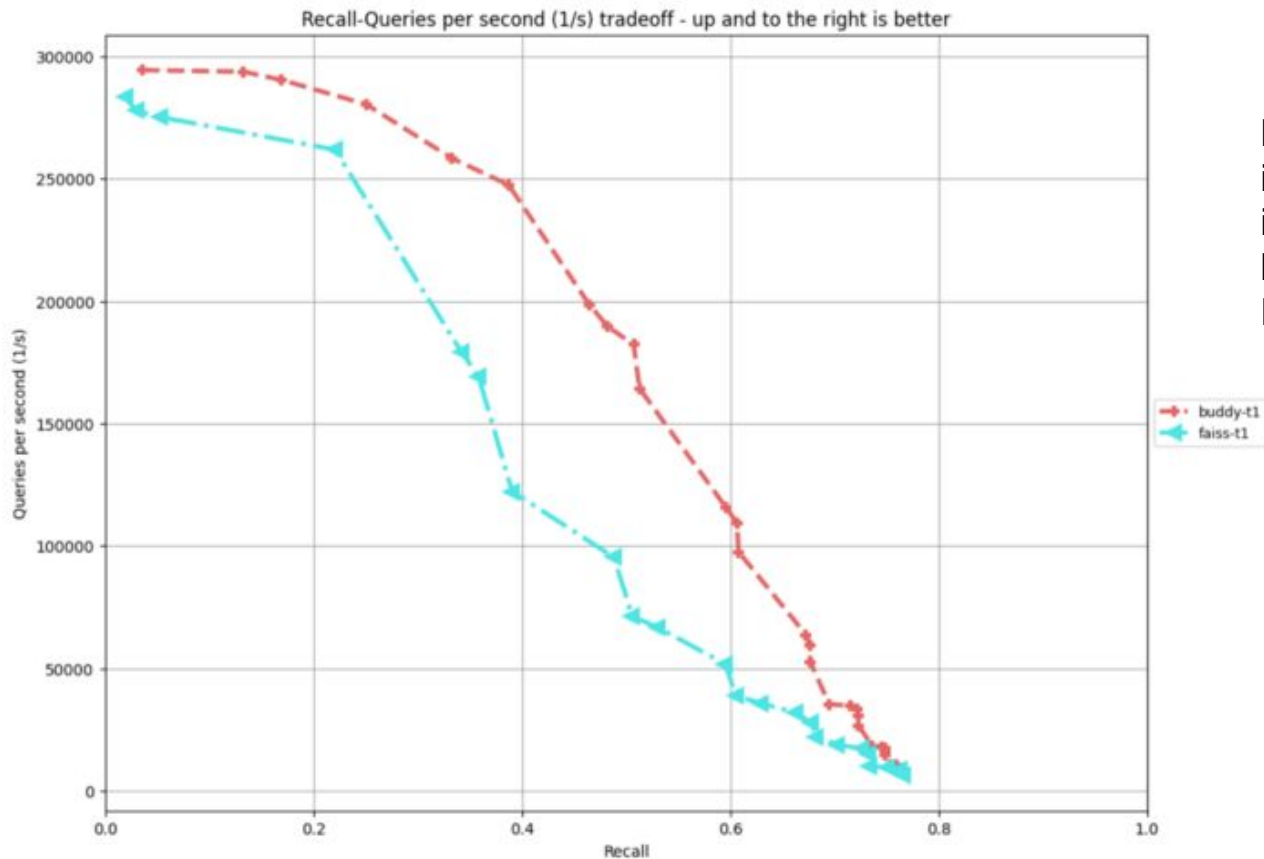
- Spotify: ANNOY
- Microsoft (Bing team): Zoom, DiskANN, SPTAG
  - Azure Cognitive Search
- Amazon: KNN based on HNSW in OpenSearch
- Google: ScaNN
- Yahoo! Japan: NGT
- Facebook: FAISS, PQ (CPU & GPU)
- Baidu: IPDG ([Baidu Cloud](#))
- Yandex
- NVidia
- Intell

# ANN algorithms



## Related topics

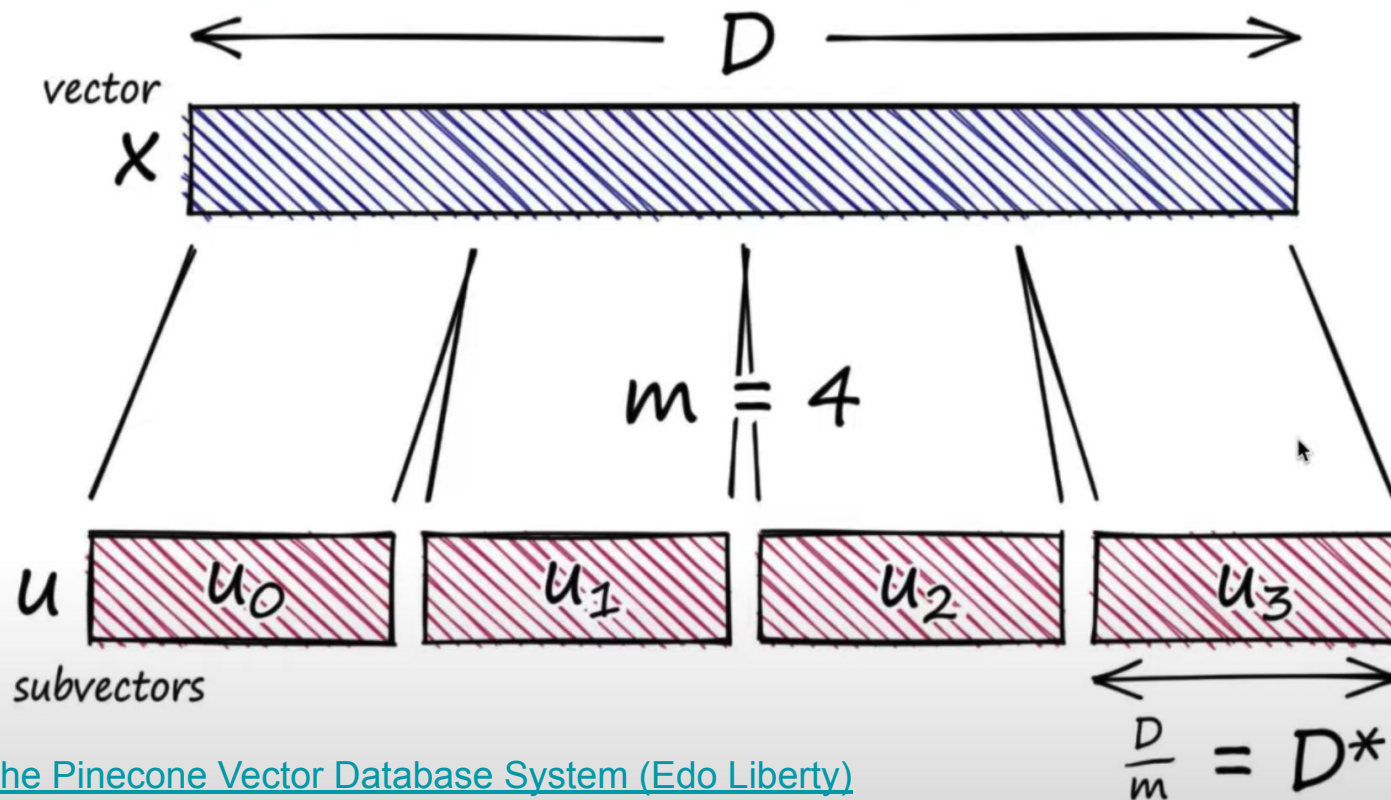




BuddyPQ  
increases **12%**  
in Recall over  
baseline  
FAISS

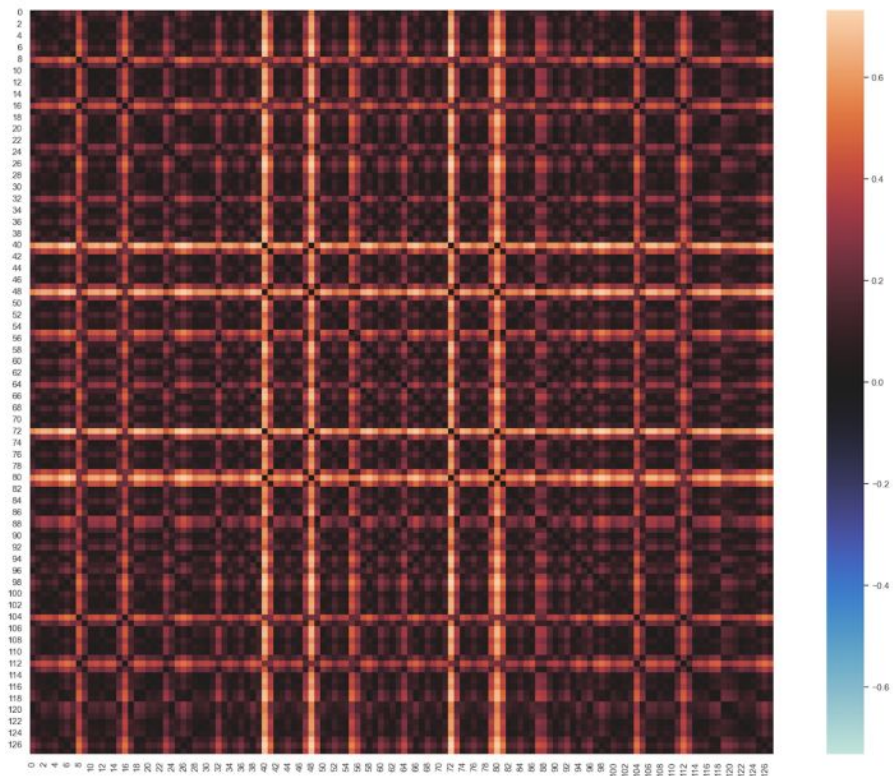
<https://bit.ly/3ApqYYQ>

# PQ (Product Quantization)

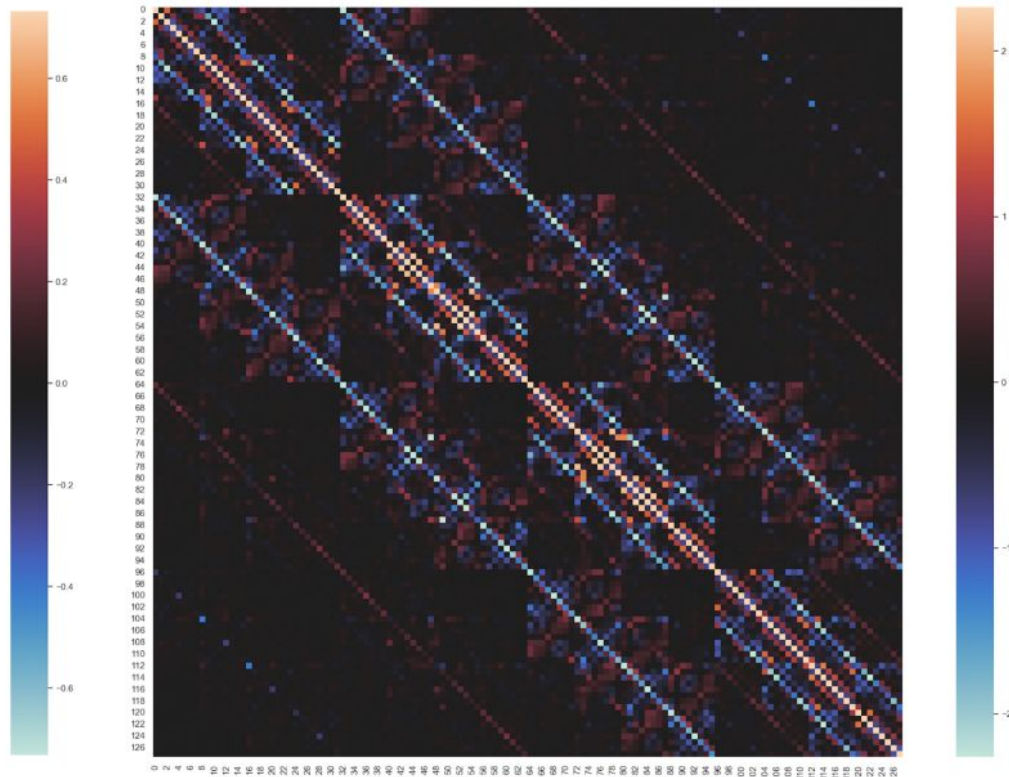


Credit: [The Pinecone Vector Database System \(Edo Liberty\)](#)

# BuddyPQ: improving over FAISS



BIGANN dataset Kolmogorov-Smirnov dimension test matrix for the first 100000 points. A higher number indicates a less similar distribution



Variance Inflation Factor (Multicollinearity) (~2.25)



# Not All Vector Databases Are Made Equal

A detailed comparison of Milvus, Pinecone, Vespa, Weaviate, Vald, GSI and Qdrant



Dmitry Kan Oct 2 · 7 min read ★



*While working on this blog post I had a privilege of interacting with all search engine key developers / leadership: Bob van Luijt and Etienne Dillocker (Weaviate), Greg Kogan (Pinecone), Pat Lasserre, George Williams (GSI Technologies Inc), Filip Haltmayer (Milvus), Jo Kristian Bergum (Vespa), Kiichiro Yukawa (Vald) and Andre Zayarni (Qdrant)*

This blog is discussed on HN: <https://news.ycombinator.com/item?id=28727816>

Update: Vector Podcast [launched!](#)



## Smaller Vector DB players: 71% are Open Source

Company	Product	Cloud	Open Source: Y/N	Algorithms
SeMI	Weaviate	Y	Y (Go)	custom HNSW
Pinecone	Pinecone	Y	N	FAISS + own
GSI	APU chip for Elasticsearch / Opensearch	N	N	Neural hashing / Hamming distance
Qdrant	Qdrant	N	Y (Rust)	HNSW (graph)
Yahoo!	Vespa	Y	Y (Java, C++)	HNSW (graph)
Ziliz	Milvus	N	Y (Go, C++, Python)	FAISS, HNSW
Yahoo!	Vald	N	Y (Go)	NGT

# Milvus

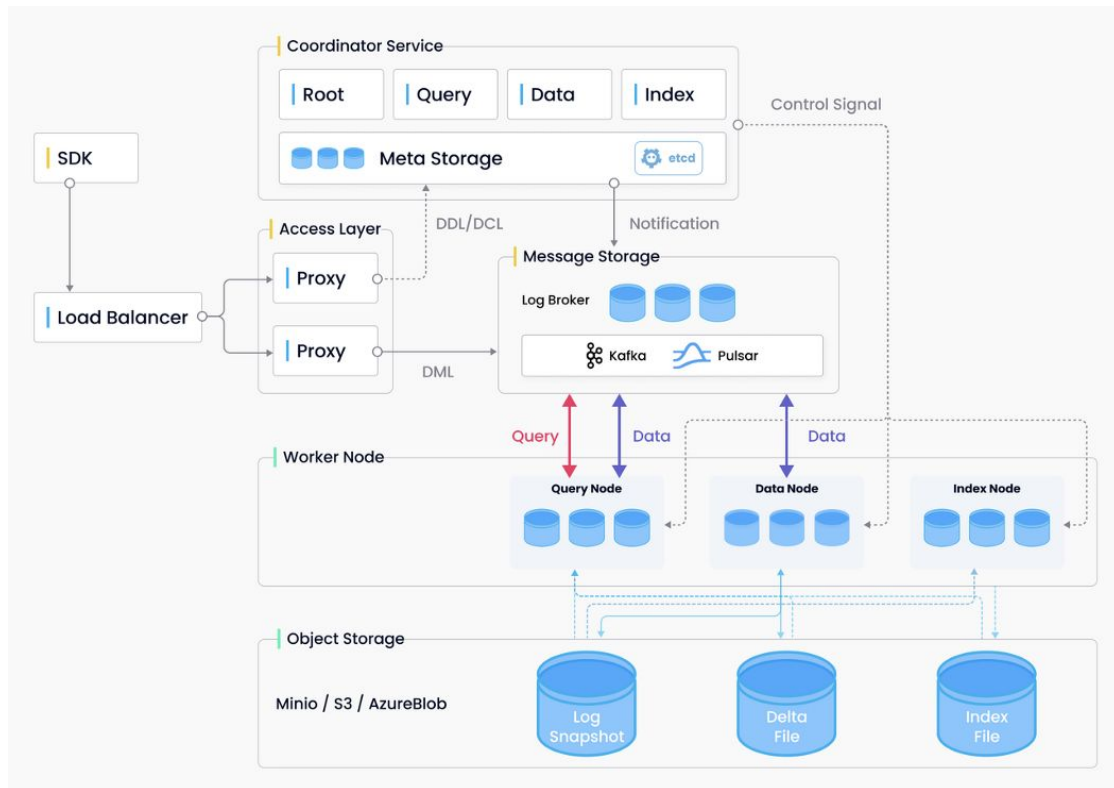
 [milvus.io](https://milvus.io)

 self-hosted vector database

 [open source](#)

Value proposition:

- attention to scalability of the entire search engine: (re)indexing and search
- ability to index data with [multiple ANN algorithms](#) to compare their performance for your use case





# Pinecone

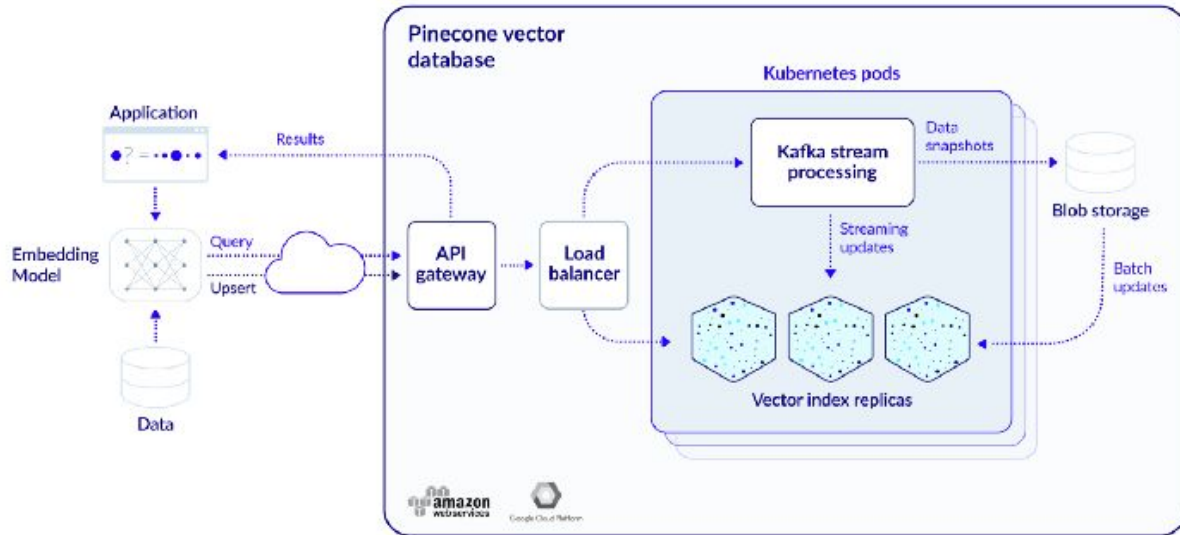
 [pinecone.io](https://pinecone.io)

 managed vector database

 close source

Value proposition:

- Fully managed vector database
- Single-stage filtering capability: search for your objects (sweaters) + filter by metadata (color, size, price) in one query





 [vespa.ai/](https://vespa.ai/)



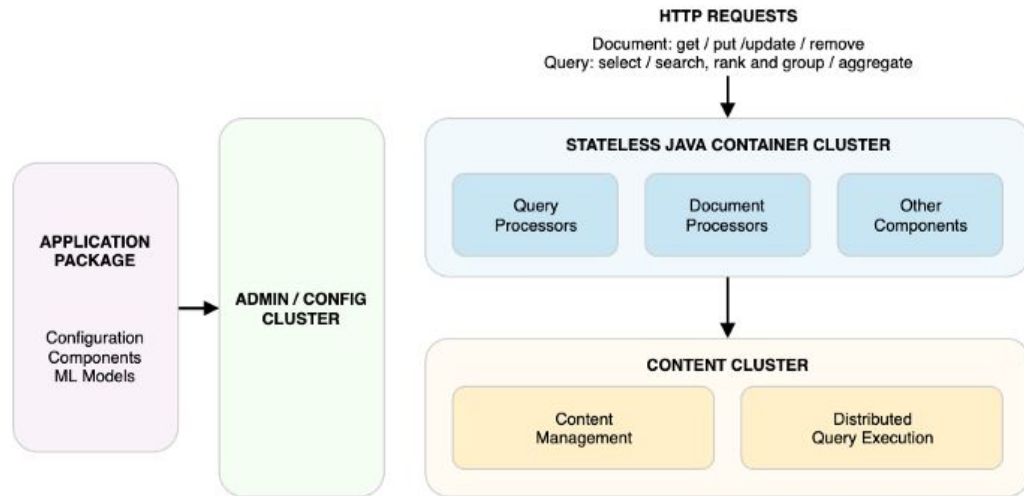
managed / self-hosted



Code: [open source](#)

Value proposition:

- low-latency computation over large data sets
- stores and indexes your data so that queries, selection and processing over the data can be performed at serving time
- customizable functionality
- deep data structures geared towards deep-learning like data science, like Tensors



# Weaviate



[semi.technology/developers/weaviate/current/](https://semi.technology/developers/weaviate/current/)



managed / self-hosted

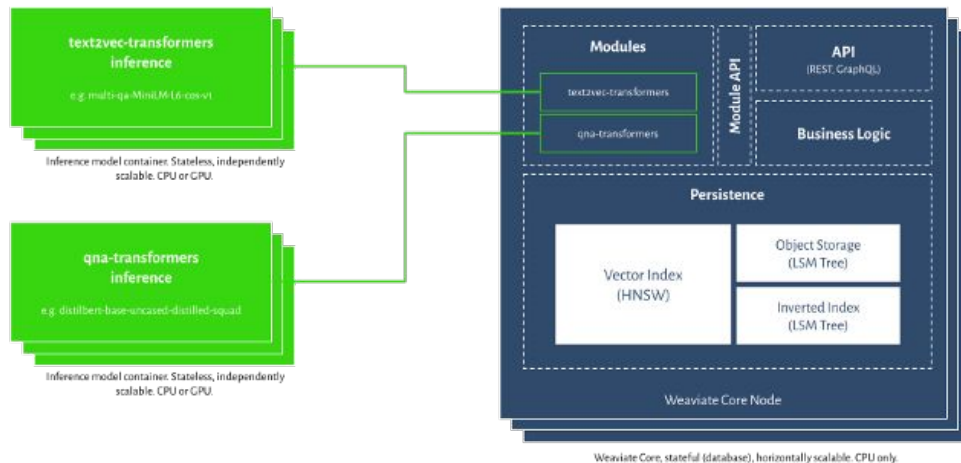


[open source](#)

Value proposition:

- Expressive query syntax
- [Graphql-like](#) interface
- combo of vector search, object storage and inverted index
- Wow-effect: Has an impressive [question answering component](#) — esp for demos

## Weaviate System Level Overview (Example with two modules)



Two modules (text2vec-transformers, qna-transformers) shown as an example. Other modules include vectorization for other media types, entity recognition, spell checking and others.

Persistence in Weaviate Core shows one shard as an example. Users can create any number of indices, each index can contain any number of shards. Shards can be distributed and/or replicated across nodes in the cluster. A shard always contains object, inverted and vector storage. Vector storage is not affected by LSM segmentation.



# Vald

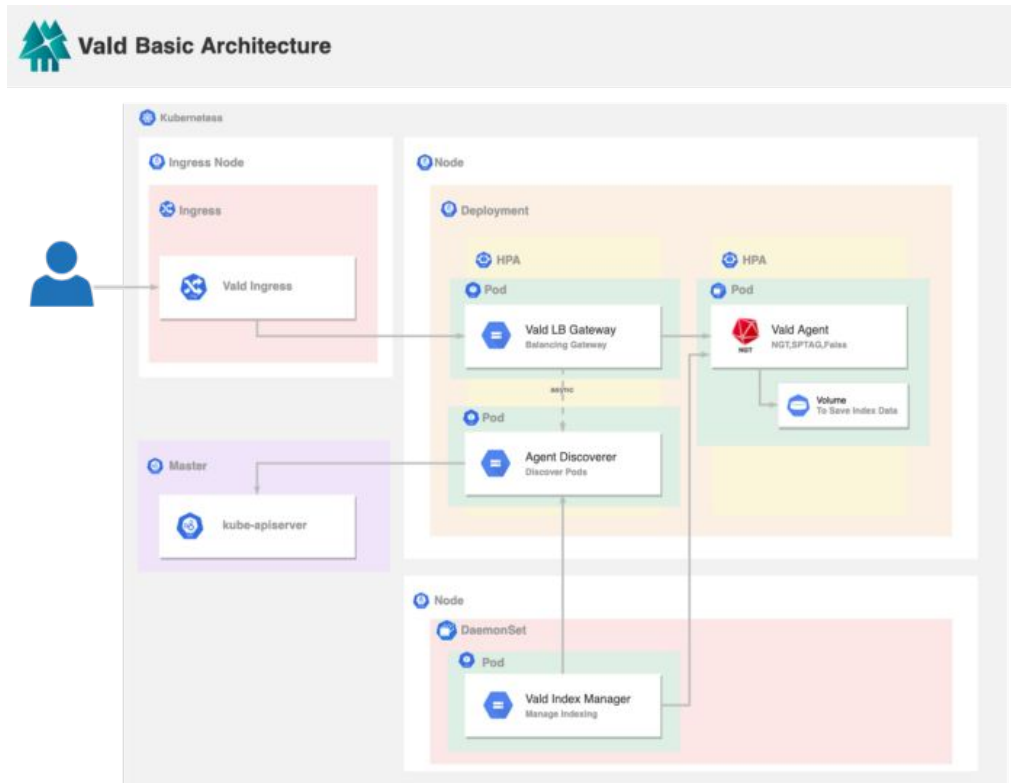
Link: [vald.vdaas.org/](https://vald.vdaas.org/)

💡 Type: Self-hosted vector database

🤖 Code: [open source](#)

Value proposition:

- Billion-scale
- Cloud-native architecture
- Fastest ANN Algo: NGT
- Custom reranking / filtering algorithm plugins



# GSI APU

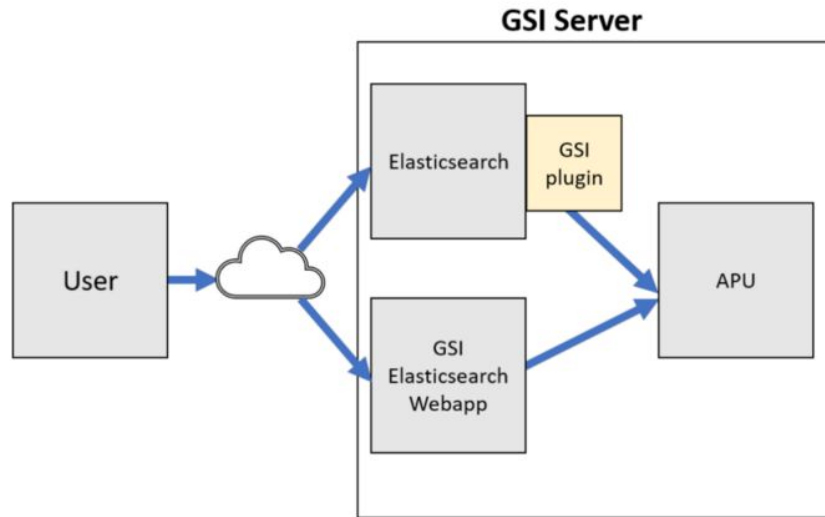
🌐 Link: [gsitechnology.com/APU](https://gsitechnology.com/APU)

💡 Type: Vector search hardware backend for your [Elasticsearch](#) / [OpenSearch](#)

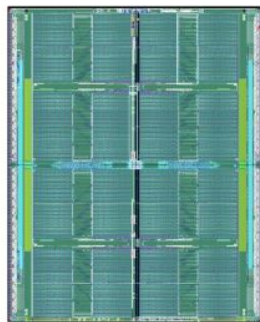
🤖 Code: close source

Value proposition:

- Billion-scale
- Extends your Elasticsearch / OpenSearch capabilities to similarity search
- On-prem / hosted APU board hosted cloud backend



## Gemini® APU Processor




- Internal Clock
  - 200 – 500 MHz
- Compute In Memory
  - 48 million 10T SRAM cells
  - 2 million units of prog "bit-logic"
- L1 Cache
  - 96Mb
- Algorithms
  - Similarity Search
  - Vector Processing
  - SAR BPA, Image Processing



# Qdrant

 [qdrant.tech/](https://qdrant.tech/)

 self-hosted vector database (Cloud in roadmap)

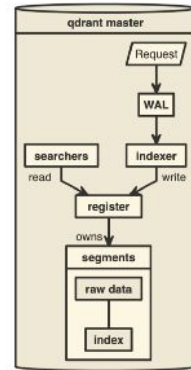
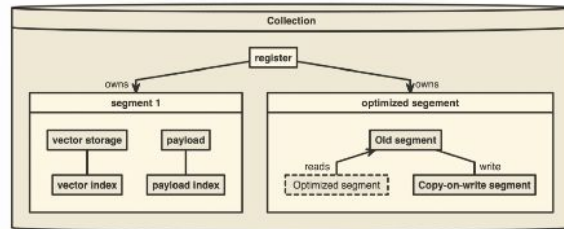
 [open source](#)

## Value proposition:

- The vector similarity engine with extended filtering support
- dynamic query planning and payload data indexing
- string matching, numerical ranges, geo-locations, and more
- [Metric Deep Learning](#)

## Qdrant Architecture

- Storage is split into **Segments**
- Segments can be re-built by the **optimizer**
- Segments are always available for search



## Semantic frameworks / layers: 57% Open Source

Company	Product	Open Source: Y/N	Focus
Deepset.ai	Haystack	Y	NLP, neural search
Jina.AI	Jina, Hub, Finetuner	Y	NLP, CV, ASR
Featureform	Feature store, EmbeddingHub	Y	All AI verticals
ZIR.AI	AI search platform	N	NLP
Hebbia.AI	Knowledge Base	N	NLP -> Finance
Rasa.ai	Virtual assistants	Y	NLP
Muves.io	Multilingual vector search	N	Multilingual search, multimodality



**user interface**

Application business logic: neural / BM25, symbolic  
filters, ranking

Multi-modal encoder / single modality encoders

Neural frameworks: Haystack, Jina.AI, ZIR.AI, Muves, Hebbia.AI, Featureform

Vector Databases: Milvus, Weaviate, Pinecone, GSI, Qdrant, Vespa, Vald, Elastiknn

KNN / ANN algorithms: HNSW, PQ, IVF, LSH, Zoom, DiskANN, BuddyPQ, ...



# How to pick a vector DB / framework

- Have own engineering?
  - Yes: go for the framework vendor / self-hosted DB
  - No: choose higher-level system, like Hebbia.AI
- Own embedding layer or OK with vector DB doing it?
  - Own: Qdrant, Milvus, Pinecone, GSI, Vespa, Vald
  - In-DB: Weaviate, Vespa
- Heavy focus on NLP?
  - YES: Consider Haystack (deepset)
  - NO: Consider Jina.AI
- Want to quickly test before investing?
  - Yes: ZIR.AI, Hebbia.AI
  - No: Jina.AI, Haystack etc

# How to pick a vector DB / framework

- Want to HOST or fine with MANAGED?
  - HOST: Vespa, Vald, Milvus, Qdrant
  - MANAGED: Pinecone, Weaviate, GSI, Hebbia.AI, ZIR.AI

# Demo: Muves

demo.muves.io

**muves**

Search products:

Miten elää onnellinen elämä

Search

Results: 5



Top product matches:



[How To Live Happy](#)

Category: Books



[How to Be Happy](#)

Category: Books



[How to Live a Prosperous Life](#)

Category: Books



[Abundance of Joy: How to Live a Joy-Filled Life](#)

Category: Books



[How Happy to Be](#)

Category: Books

## Trends in ML at large

- Model hubs (e.g. Hugging Face) → ML community shares progress quickly (similar to what GitHub did to sharing code)
- Deep Learning → multimodal: CLIP (text from images), DALL-E (images from text)
- MLOps optimize experimentation and deployments: determined.ai, DVC, MLflow / Kubeflow

# Get practical

- Code: <https://github.com/DmitryKey/bert-solr-search>
- Supported Engines: Solr, Elasticsearch, OpenSearch, GSI, [hnswlib]
- Supported LMs: BERT, SBERT, [theoretically any]

# Q&A demo

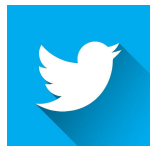
Questions:

1. Why did Peloton shares fall?
2. How are articles created on TechCrunch?
3. What is ethical AI?

<https://techcrunch.vectors.network/>

[Weaviate console](#)

Thank you! ❤️



[twitter.com/DmitryKan](https://twitter.com/DmitryKan)



[youtube.com/c/VectorPodcast](https://youtube.com/c/VectorPodcast)



<https://spoti.fi/3sRXcdn>

# Links

1. Martin Fowler's talk on NoSQL databases: [https://www.youtube.com/watch?v=qI\\_g07C\\_Q5I](https://www.youtube.com/watch?v=qI_g07C_Q5I)
2. Neural search vs Inverted search  
<https://trends.google.com/trends/explore?date=all&q=neural%20search.inverted%20search>
3. Not All Vector Databases Are Made Equal  
<https://towardsdatascience.com/milvus-pinecone-vespa-weaviate-vald-gsi-what-unites-these-buzz-words-and-what-makes-each-9c65a3bd0696>
4. HN thread: <https://news.ycombinator.com/item?id=28727816>
5. A survey of PQ and related methods: <https://faiss.ai/>
6. Vector Podcast on YouTube:  
<https://www.youtube.com/channel/UCCIMPfR7TXyDvIDRXjVhP1g>
7. Vector Podcast on Spotify: <https://open.spotify.com/show/13JO3vhMf7nAqcpvllgOY6>
8. Vector Podcast on Apple Podcasts:  
<https://podcasts.apple.com/us/podcast/vector-podcast/id1587568733>
9. BERT, Solr, Elasticsearch, OpenSearch, HNSWlib – in Python:  
<https://github.com/DmitryKey/bert-solr-search>