Results of the Big ANN: NeurIPS'23 Competition

Harsha Vardhan Simhadri (Microsoft) Martin Aumüller (IT U. Copenhagen) Amir Ingber (Pinecone) Matthijs Douze (Meta AI) George Williams Dmitry Baranchuk (Yandex) Edo Liberty (Pinecone) Frank Liu (Zilliz)

Participating Teams: Shanghai Jiao Tong University, Fudan University, Baidu, University of Maryland. and Carnegie Mellon University.

Introduction

Approximate Nearest Neighbor (ANN) search is critical for LLMs (RAG), computer vision, and recommendation systems. While previous challenges focused on scaling standard dense vector indexing, the 2023 Big ANN Challenge[1] addressed **practical, complex variants** of ANN search encountered in real-world applications.

Key Goals

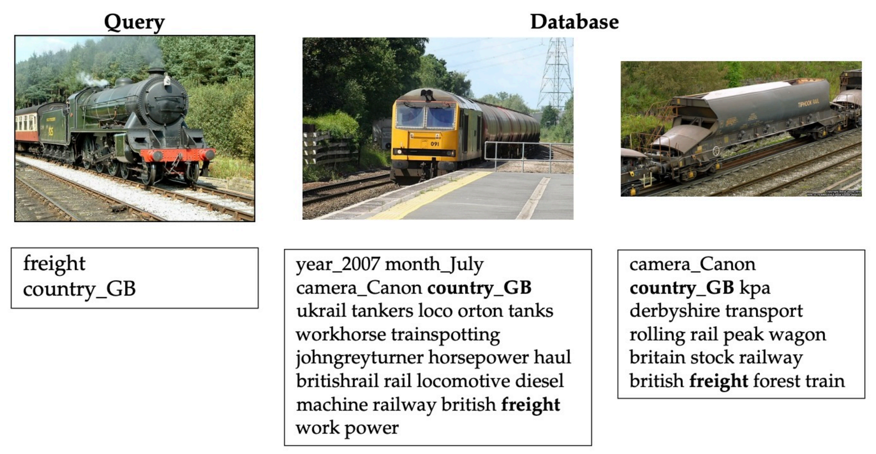
- Move beyond standard dense indexing.
- Address diverse data distributions and types.
- Evaluate on constrained hardware (16GB RAM).
- Promote open-source contributions.

Tracks & Datasets

1. Filtered Search

Dataset: YFCC 100M[2] (CLIP embeddings + Tags)

Search for nearest neighbors that *also* match specific metadata tags.



2. Out-Of-Distribution (OOD)

Dataset: Yandex Text-to-Image 10M [4]

Database (Images) and Query (Text) vectors have different distributions in the shared space. Standard indices often fail to provide high recall.

3. Sparse Search

Dataset: MSMARCO[3] (SPLADE model)

High-dimensional vectors (>30k dim) with few non-zero elements (~120). Optimized for inverted indices and specialized graphs.

4. Streaming

Dataset: MS Turing[4] (10M subset)

Indices must handle a "runbook" of Insertions, Deletions, and Searches under strict memory (8GB) and time limits.

Evaluation Protocol

Hardware	Metric
Azure D8lds_v5	10-recall@10
8 vCPUs, 16GB RAM	& QPS (Throughput)

Ranking based on highest throughput (QPS) achieving at least 90% recall. Streaming track used highest recall across searches.

Results & Winners

Filtered Track — Winner: ParlayANN

- **ParlayANN** 11x baseline speed using inverted index for tags + Vamana graph for dense vectors.

Filter Track: All Results

OOD Track — Winners: MysteryANN & PyANNS

- **MysteryANN:** Bipartite graph between base and query samples to adapt the index.
- **PyANNS:** Relied on highly optimized Vamana graph + quantization.

Algorithm	QPS
pyanns	22296
mysteryann-dif	22492
sustech-ood	13772
puck	8700
vamana	6753
ngt	6374
epsearch	5877
cufe	3561

OOD Track: All Results

Sparse Track — Winners: PyANNS & GrassRMA

- **PyANNS:** Quantized HNSW. Refinement step used full vectors to recover accuracy.
- **GrassRMA:** Optimized memory access patterns for sparse data in graphs.

Algorithm	QPS (private)	QPS (public)
pyanns	6500	8732
shnsw	5078	7137
nle	1313	2359
sustech-whu	788	1015
cufe	98	105
linscan	95	93

Sparse Track: All Results

Streaming Track — Winner: PyANNS

- **PyANNS:** Used DiskANN[4] strategy with 8-bit scalar quantization. Quantization allowed deeper graph search within the time limit, maintaining high recall despite deletions.

Algorithm	Recall
pyanns	0.8865
hwtl_sdu_anns_stream	0.7693
diskann	0.7218
cufe	0.6481
puck	0.0921

Streaming Track: All Results

Discussion & Future Work

The 2023 challenge demonstrated a significant leap in performance over industry baselines, driven by specialized data structures rather than raw scaling.

Key Trends

- **Graph Indices** are dominant, even for sparse/OOD.
- **Quantization** is essential for efficiency on constrained hardware.
- **Hybrid Indices** (Graph + Inverted) needed for Filtered search.

Impact

- Established benchmarks for "ragged" real-world data.
- Highlighted need for "deletions" support in streaming indices.

Github Repository

We open-sourced the competition evaluation framework and all the participating algorithms. Scan the QR Code and get detailed analysis, access to the algorithms, and more information about getting involved with future competitions!

References:

[1] Simhadri et al. "Results of the NeurIPS'21 Challenge on Billion-Sclae Approximate Nearest Neighbor Search." NeurIPS Competition and Demos, PMLR, 2021.

[2] Thomee et al. "YFCC100M: The New Data in Multimedia Research" Comm. ACM, 2016.

[3] Nguyen et al. "MS MARCO: A Human Generated Machine Reading Comprehension Dataset.", 2016.

[4] Simhadri et al. "DiskANN: Graph-structured Indices for Scalable, Fast, Fresh and Filtered ANN Search", https://github.com/microsoft/DiskANN, 2023.