

# OAK

Cedric Sirianni, Mithi Jethwa, Lachlan Kermode

## ACM Reference Format:

Cedric Sirianni, Mithi Jethwa, Lachlan Kermode. 2024. OAK. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 1 page. <https://doi.org/10.1145/nnnnnnn.nnnnn>

## ABSTRACT

Vector databases are growing in popularity as they become widely used in similarity search and RAG systems as part of ML workloads. At the same time, applications increasingly leverage mixed-modality data, requiring support for search over *vector data* such as images and text, and *structured data* such as metadata and keywords, simultaneously. Recent work in ACORN helps improve the feasibility of this *hybrid search* by providing a performant and predicate-agnostic index built on Hierarchical Navigable Small Worlds (HNSW), a state-of-the-art graph based index for approximate nearest neighbor search (ANNS). However, ACORN does not take into consideration predicate access patterns, leaving room for performance improvement under certain modal workloads. To address this, we present OAK, a system for creating predicate subgraphs. To evaluate OAK, we compare OAK to ACORN on .... We show that OAK achieves improved performance ... Our code is available at: <https://github.com/breezykermode/oak>.

## INTRODUCTION

## IMPLEMENTATION

## EVALUATION

DEEP1B [1] and SIFT1B [2] are datasets commonly used to test performance and accuracy for VectorDBs. Similarly, the big ANN benchmarks repository [3, 4] provides various datasets calibrated to four different classes of load: filtered (including metadata), out-of distribution (queries are significantly different in distribution than the database), sparse (vectors have a majority of zero values), and streaming (load includes insertion and deletion operations).

## BIBLIOGRAPHY

- [1] Artem Babenko and Victor Lempitsky. 2016. Efficient Indexing of Billion-Scale Datasets of Deep Descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2055–2063.
- [2] Hervé Jégou, Romain Tavenard, Matthijs Douze, and Laurent Amsaleg. 2011. Searching in one billion vectors: Re-rank with source coding. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011. 861–864. <https://doi.org/10.1109/ICASSP.2011.5946540>
- [3] Harsha Vardhan Simhadri. 2024. Harsha-simhadri/big-ann-benchmarks.
- [4] Harsha Vardhan Simhadri, Martin Aumüller, Amir Ingber, Matthijs Douze, George Williams, Magdalen Dobson Manohar, Dmitry Baranchuk, Edo Liberty, Frank Liu, Ben Landrum, Mazin Karjekar, Laxman Dhulipala, Meng Chen, Yue Chen, Rui Ma, Kai Zhang, Yuzheng Cai, Jiayang Shi, Yizhuo Chen, Weiguo Zheng, Zihao Wan, Jie Yin, and Ben Huang. 2024. Results of the Big ANN: NeurIPS'23 competition. <https://doi.org/10.48550/arXiv.2409.17424>

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*Conference'17, July 2017, Washington, DC, USA*

© 2024 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>