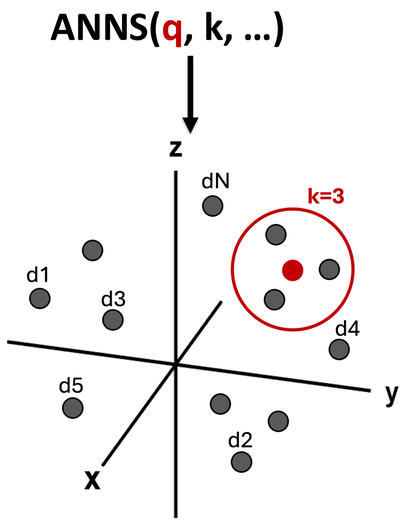
DARTH: Declarative Recall Through Early Termination for ANN Search

Manos Chatzakis[1], Yannis Papakonstantinou[2], Themis Palpanas[1]
Université Paris Cité, LIPADE[1], Google[2]
manos.chatzaki@gmail.com

Motivation

- Nearest Neighbor Vector Search (NNS) is the task of searching for the **top-k nearest vectors** of a given **query vector** in a high-dimensional space
- Approximate NNS **(ANNS)** relaxes NNS by enabling error in returned results to significantly speed up the search
- ANNS is the **backbone of various data management tasks**, such as Information Retrieval, Recommender Systems, RAG



Current ANNS Interface

- ANNS is applied using vector indexes that enable faster search (e.g., **k-NN graph, IVF**)
- Approximate vector search algorithms support ANNS with an interface of the form:

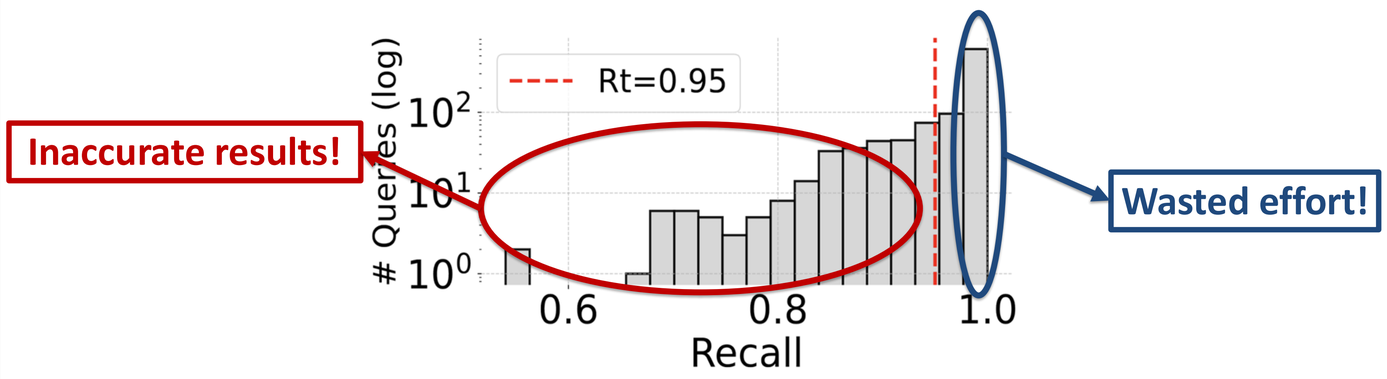
ANNS(q, k, *params)

where *params are the search parameters (e.g., **efSearch**)

- Each ANNS algorithm has its own tunable parameters
- Fundamental recall vs latency tradeoff which is configured by *params
- ANNS users have to **tune parameters via cumbersome and time consuming trial and error to achieve target recall**

Problems

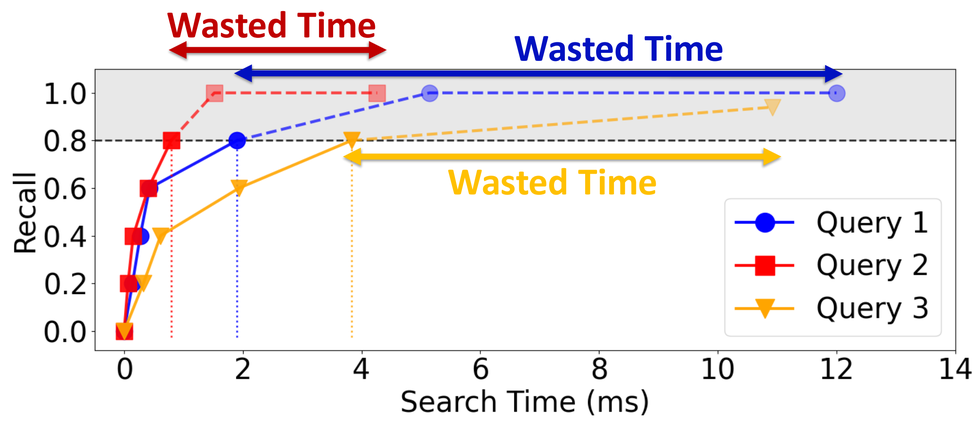
- Parameter tuning is inadequate: suboptimal results for individual queries
  - Hard queries **undershoot** (low recall) → **inaccurate results**
  - Easy queries **overshoot** (slow processing) → **wasted effort, monetary costs**



Tuning is expensive and leads to suboptimal results for individual queries!

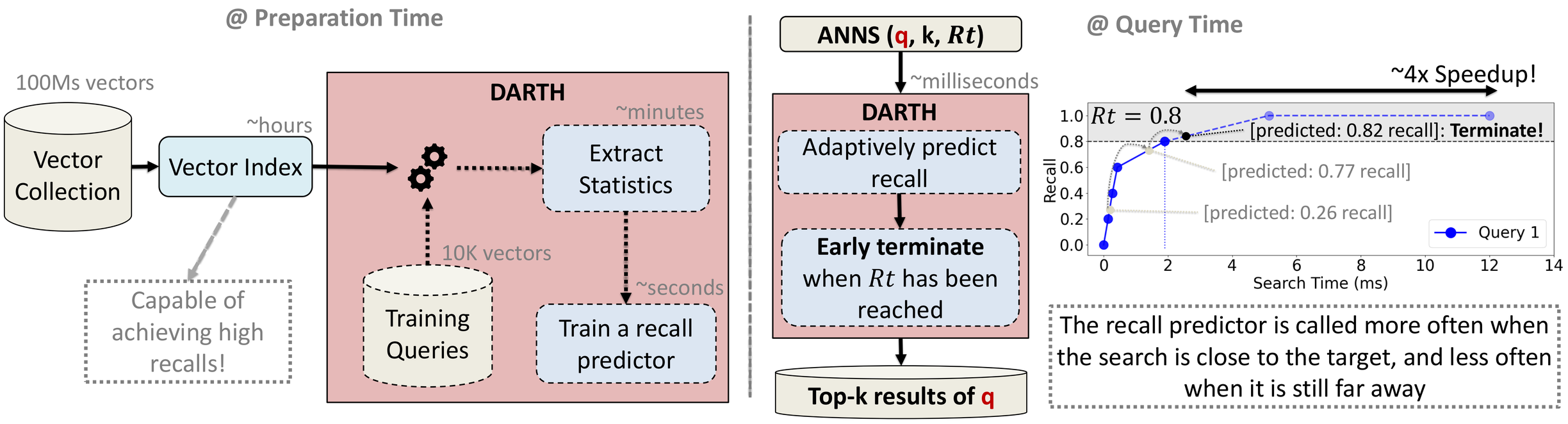
Key Observation

A query reaching high recall can serve all lower recall levels
→ **Early termination opportunity!**



Different queries reach different recalls at different times!

The DARTH Approach: Predict the recall!

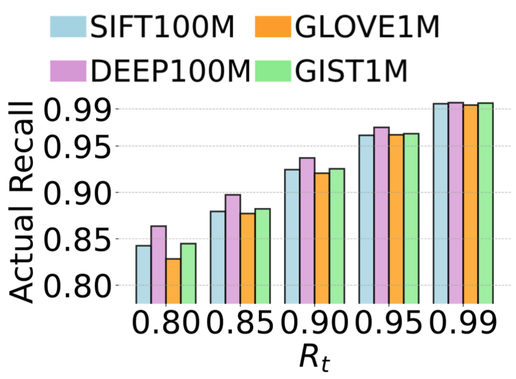


Natural support for any recall target without explicit tuning!

Experimental Evaluation

Datasets: ***SIFT100M, DEEP100M, GLOVE1M, GIST1M, T2I100M****; 1000 queries*

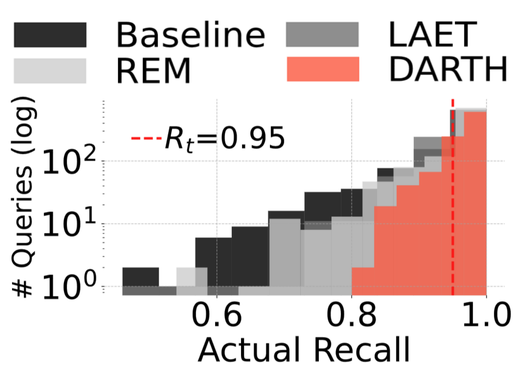
Recall Performance



- Satisfies requested recall targets in all configurations and datasets
- Speedup of up to 15x compared to plain search without early termination

Achieves all recall targets, while achieving speedups up to 15x!

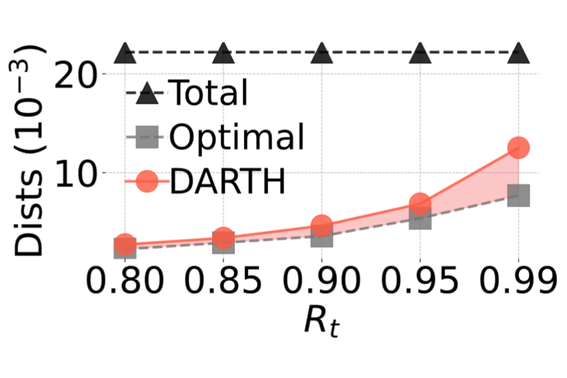
Recall Distribution



- DARTH has only ~13% of queries under target, worst recall is ~0.80
- Best competitor has ~21% of queries under target, worst recall is ~0.58

Superior recall distribution!

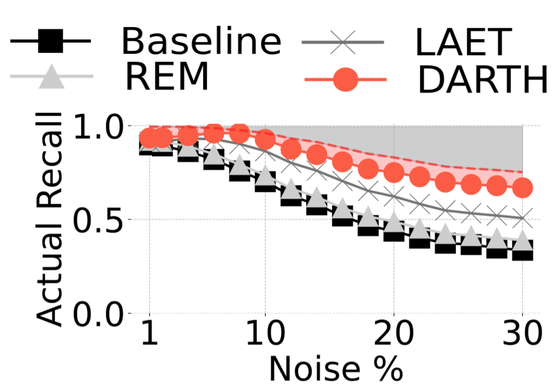
Termination Points



- Near-optimal termination points by DARTH: only 5% more distance calculations than the optimal
- Optimal is infeasible in practice

Near-optimal early termination points!

Hard Queries



- We create hard queries by adding noise
- The red line indicates the maximum recall attained by the index
- DARTH is the only method resilient to hard queries

Resilience to hard queries!