UM6P | College of Computing

Graph-Based Vector Search
Recent Advances & Future Directions

Prof. Karima Echihabi

UM6P College of Computing, Morocco

Talk at the University of Waterloo, Canada

July 28th, 2025

Introduction

Research

Main Topics

Main Topics

- Scalable and responsible data science
  - Similarity search
    - Data structures/algorithms
    - Representation learning
  - Data science
    - GenAI pipelines (e.g., RAG)
    - Data valuation
    - Health
- Computer science education
  - Survey on leading universities across the Arab region
  - Study of CS programs in the African/Arab region
  - Booklet on best practices for supporting transitions from PhD to Professorship

Research
Collaborators

Research Collaborators

Topic	Collaborators	Affiliation
Vector Search/ Data Mining	Damien Hilloulin, Marco Arnaboldi, Ioannis Alagiannis, Vlad Haprian Themis Palpanas Theophanis Tsandilas Anastasia Bezerianos Panagiota Fatourou	Oracle Labs, Zurich Paris Cite, France Inria, France Paris Saclay, France Univ. Crete, Greece
Data Valuation	Mardavij Roozbehani, Thibault Horel, Munther Dahleh	MIT, USA
Health Analytics	Yousef Yeganeh, Azade Farshad, Nassir Navab Amal Fadaili Anis Hasnaoui Gbenga Peter Oderinde, Stephen Peter Akpulu Kun-Hsing Yu	TUM, Germany Amana Pathology Lab, Morocco Faculty of Medicine, Tunisia Ahmadu Bello Univ., Nigeria Harvard Medical School, USA
Computer Science Education	Sherif G. Aly, Seif Eldawlatly Slim Abdennadher Khaled Shuaib Joe Tekli	AUC, Egypt GUC, Egypt UAE Univ., UAE Lebanese American Univ, Lebanon

Research
Team

Team

Name	Hasnae Zerouaoui
Position	Postdoctoral Fellow
Awards/ Employment	Microsoft Research PhD Fellowship NVIDIA Ambassador
Project	Representation Learning & Health Analytics

Team

Name	Hasnae Zerouaoui	Ilias Azizi
Position	Postdoctoral Fellow	Alumni PhD Student
Awards/ Employment	Microsoft Research PhD Fellowship NVIDIA Ambassador	amazon Research Internship cnrs PostDoc 2 VLDB Grants
Project	Representation Learning & Health Analytics	Approximate Graph-Based Vector Search

Team



Name	Hasnae Zerouaoui	Ilias Azizi	Khaoula Abdenouri
Position	Postdoctoral Fellow	Alumni PhD Student	3rd year PhD Student
Awards/ Employment	Microsoft Research PhD Fellowship NVIDIA Ambassador	amazon Research Internship cnrs PostDoc 2 VLDB Grants	
Project	Representation Learning & Health Analytics	Approximate Graph-Based Vector Search	Exact Tree-Based Vector Search

Team

Name	Hasnae Zerouaoui	Ilias Azizi	Khaoula Abdenouri	Anas Ait Aomar
Position	Postdoctoral Fellow	Alumni PhD Student	3rd year PhD Student	2nd year PhD Student
Awards/ Employment	Microsoft Research PhD Fellowship NVIDIA Ambassador	amazon Research Internship cnrs PostDoc 2 VLDB Grants		ORACLE Research PhD Fellowship UNIVERSITY OF WATERLOO SIGMOD Grant
Project	Representation Learning & Health Analytics	Approximate Graph-Based Vector Search	Exact Tree-Based Vector Search	Approximate Graph-Based Hybrid Vector Search

Team

Name	Hasnae Zerouaoui	Ilias Azizi	Khaoula Abdenouri	Anas Ait Aomar	Firdawse Guerbouzi
Position	Postdoctoral Fellow	Alumni PhD Student	3rd year PhD Student	2nd year PhD Student	1st year PhD Student
Awards/ Employment	Microsoft Research PhD Fellowship NVIDIA Ambassador	amazon Research Internship CNRS PostDoc 2 VLDB Grants		ORACLE Research PhD Fellowship University of Waterloo SIGMOD Grant	NVIDIA Research Associate MICCAI Travel Grant
Project	Representation Learning & Health Analytics	Approximate Graph-Based Vector Search	Exact Tree-Based Vector Search	Approximate Graph-Based Hybrid Vector Search	Representation Learning & Health Analytics

Team

Name	Hasnae Zerouaoui	Ilias Azizi	Khaoula Abdenouri	Anas Ait Aomar	Firdawse Guerbouzi	Mehdi Touil
Position	Postdoctoral Fellow	Alumni PhD Student	3rd year PhD Student	2nd year PhD Student	1st year PhD Student	1st year PhD Student
Awards/ Employment	Microsoft Research PhD Fellowship NVIDIA Ambassador	amazon Research Internship CNRS PostDoc 2 VLDB Grants		ORACLE Research PhD Fellowship UNIVERSITY OF WATERLOO SIGMOD Grant	NVIDIA Research Associate MICCAI Travel Grant	MIT
Project	Representation Learning & Health Analytics	Approximate Graph-Based Vector Search	Exact Tree-Based Vector Search	Approximate Graph-Based Hybrid Vector Search	Representation Learning & Health Analytics	Data Valuation

Team

Name	Hasnae Zerouaoui	Ilias Azizi	Khaoula Abdenouri	Anas Ait Aomar	Firdawse Guerbouzi	Mehdi Touil	Reda Lefdali	Abdelatif Bouzid
Position	Postdoctoral Fellow	Alumni PhD Student	3rd year PhD Student	2nd year PhD Student	1st year PhD Student	1st year PhD Student	Data Scientist	Data Engineer
Awards/ Employment	Microsoft Research PhD Fellowship NVIDIA Ambassador	amazon Research Internship cnrs PostDoc 2 VLDB Grants		ORACLE Research PhD Fellowship UNIVERSITY OF WATERLOO SIGMOD Grant	NVIDIA Research Associate MICCAI Travel Grant	MIT	OCP	OCP
Project	Representation Learning & Health Analytics	Approximate Graph-Based Vector Search	Exact Tree-Based Vector Search	Approximate Graph-Based Hybrid Vector Search	Representation Learning & Health Analytics	Data Valuation	Health Analytics/ Data Mining	RAG

Vector Search

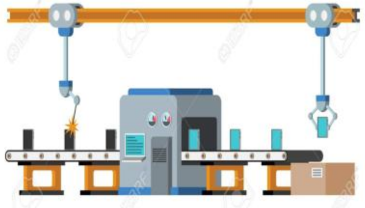
Vector Search

Overview

High-D Data is Everywhere


Finance


Paleontology

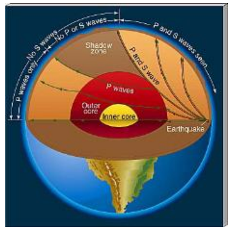

Manufacturing

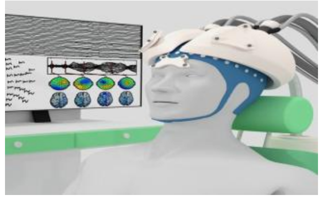

Aviation


Agriculture

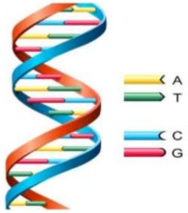

Astronomy

Criminology


Seismology


Neuroscience


Medicine



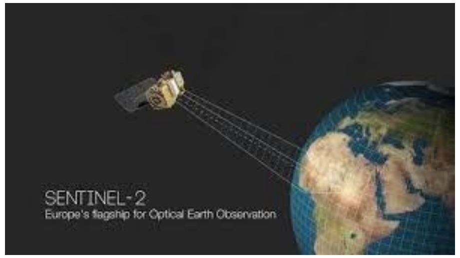
Biology

High-D Collections are Massive

≈ 500 ZB per year

≈ 130 TB

> 40 PB per day



> 5 TB per day

> 500 TB per day

1 PB = 1 thousand TB

1 ZB = 1 billion TB

1 PB = 1 thousand TB = 10^{15} bytes

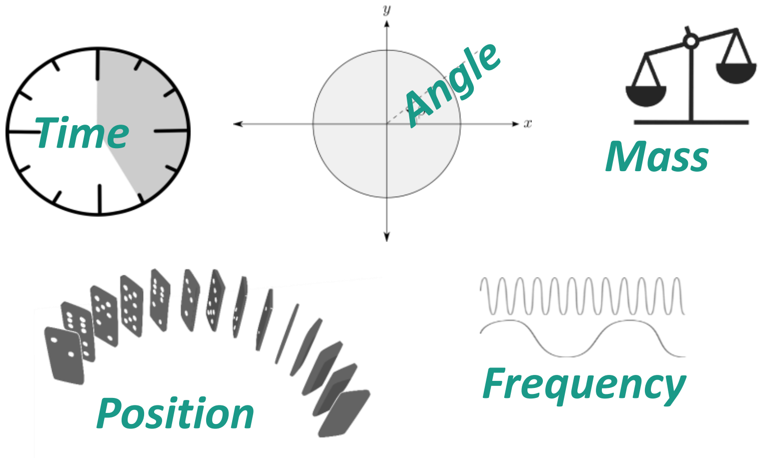
1 ZB = 1 billion TB = =10^{21} bytes

Popular High-D Data

Popular High-D Data

Data Series

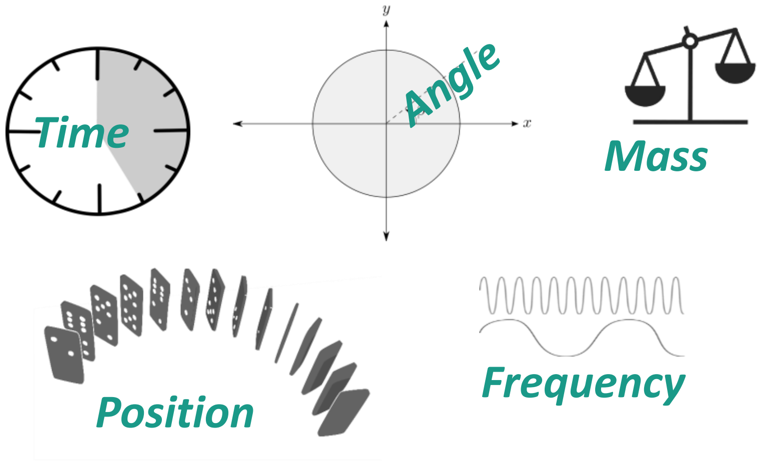
A collection of points ordered over a dimension



Popular High-D Data

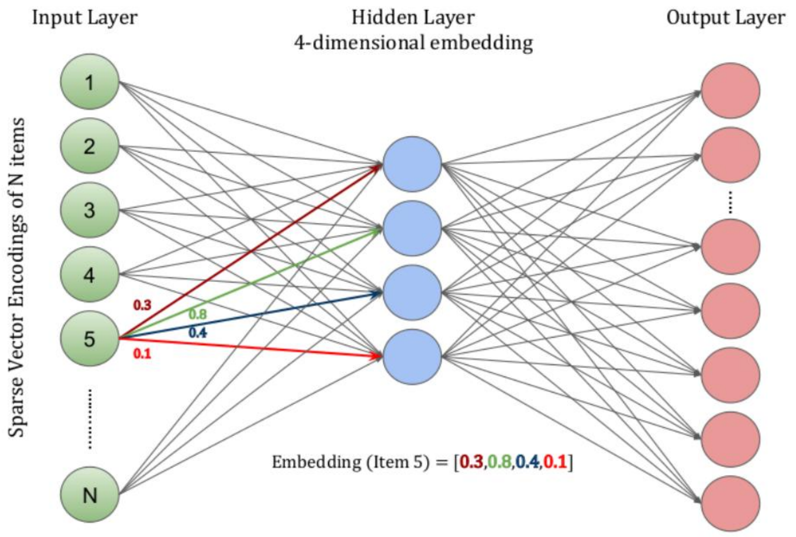
Data Series

A collection of points ordered over a dimension



Deep Embeddings

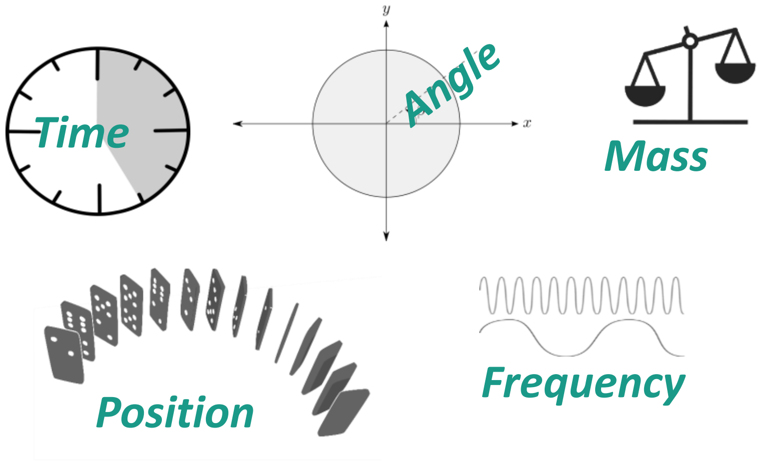
A feature vector learned from data using a DNN



Popular High-D Data

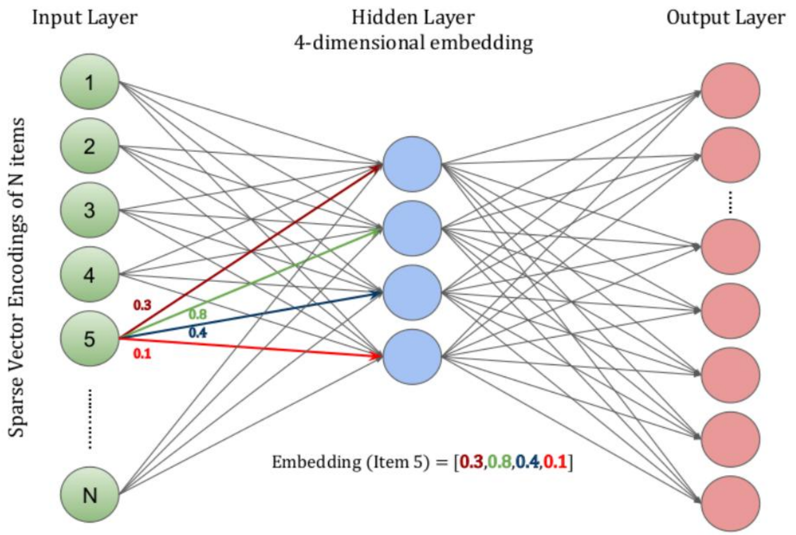
Data Series

A collection of points ordered over a dimension



Deep Embeddings

A feature vector learned from data using a DNN

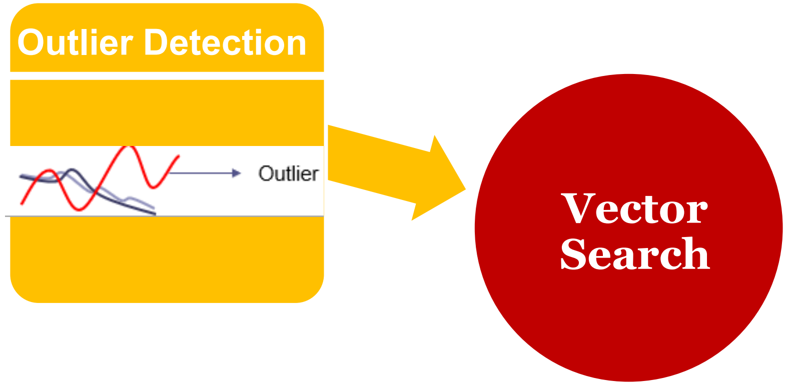


embedded
text, images, video, graphs, etc.

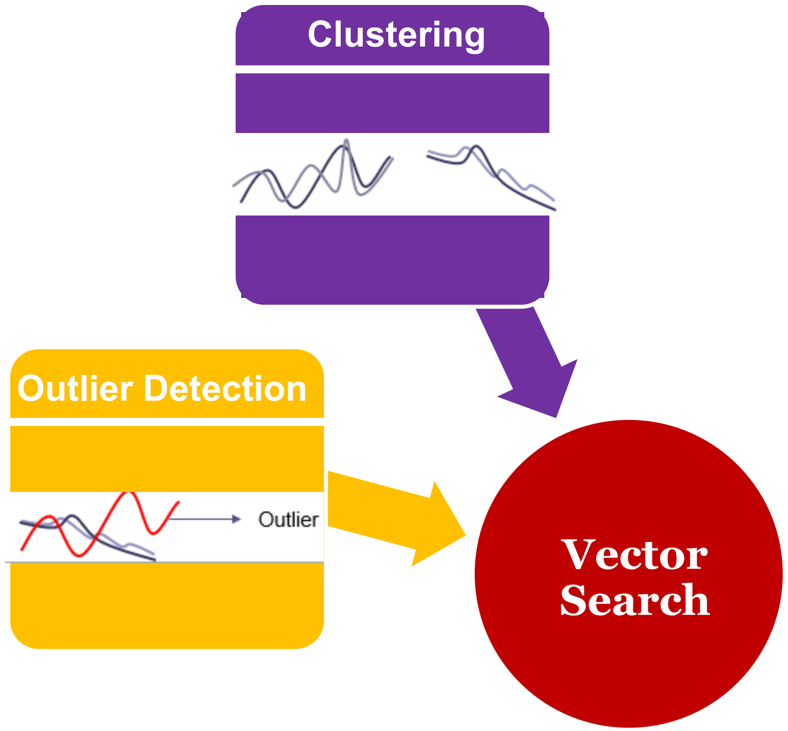
Vector Similarity Search at the Core of Data Science

Vector Search

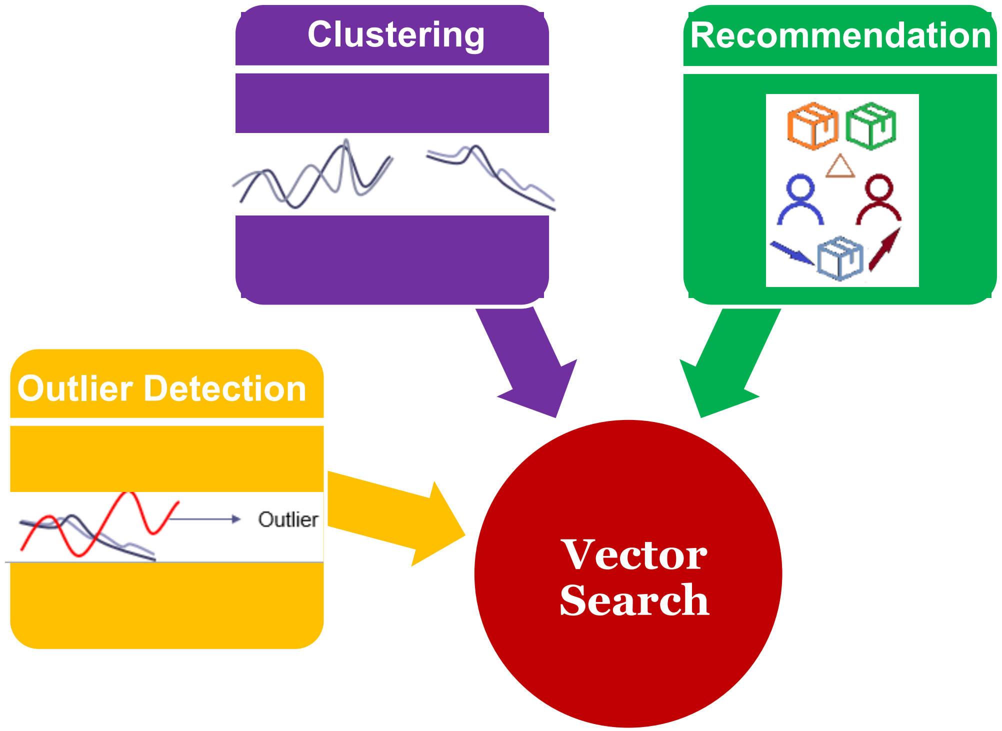
Vector Similarity Search at the Core of Data Science



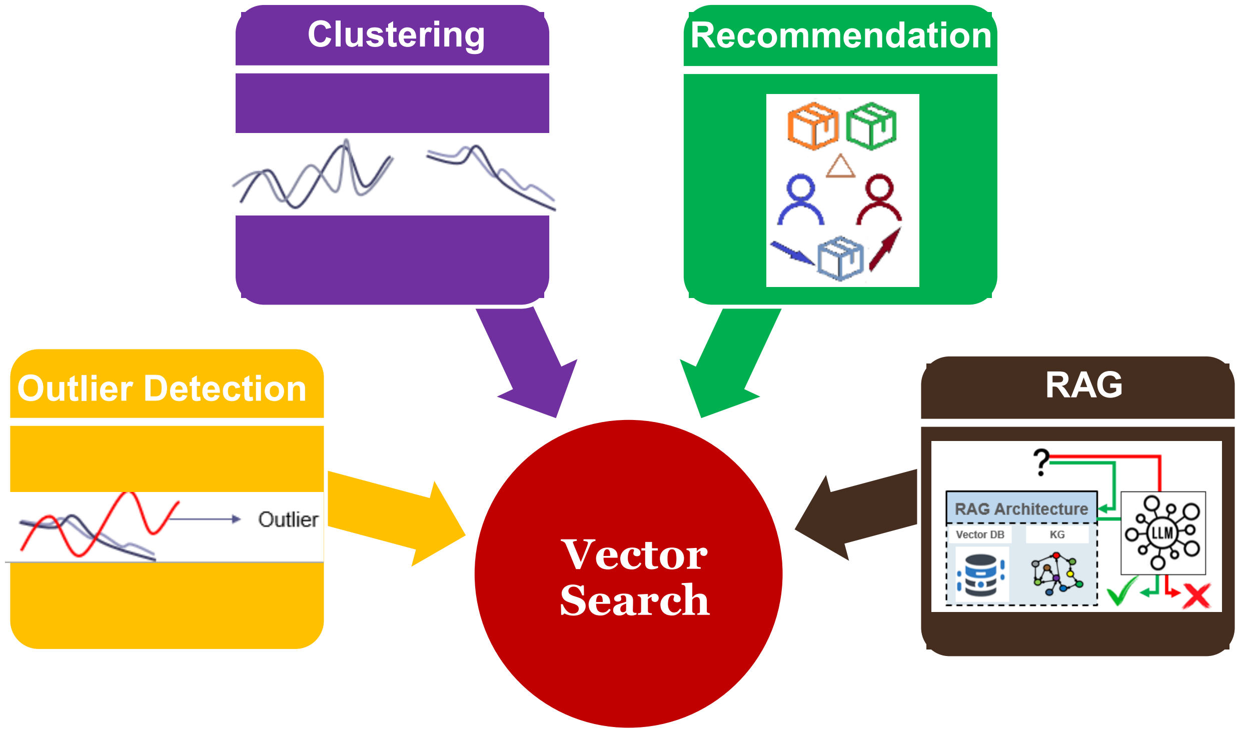
Vector Similarity Search at the Core of Data Science



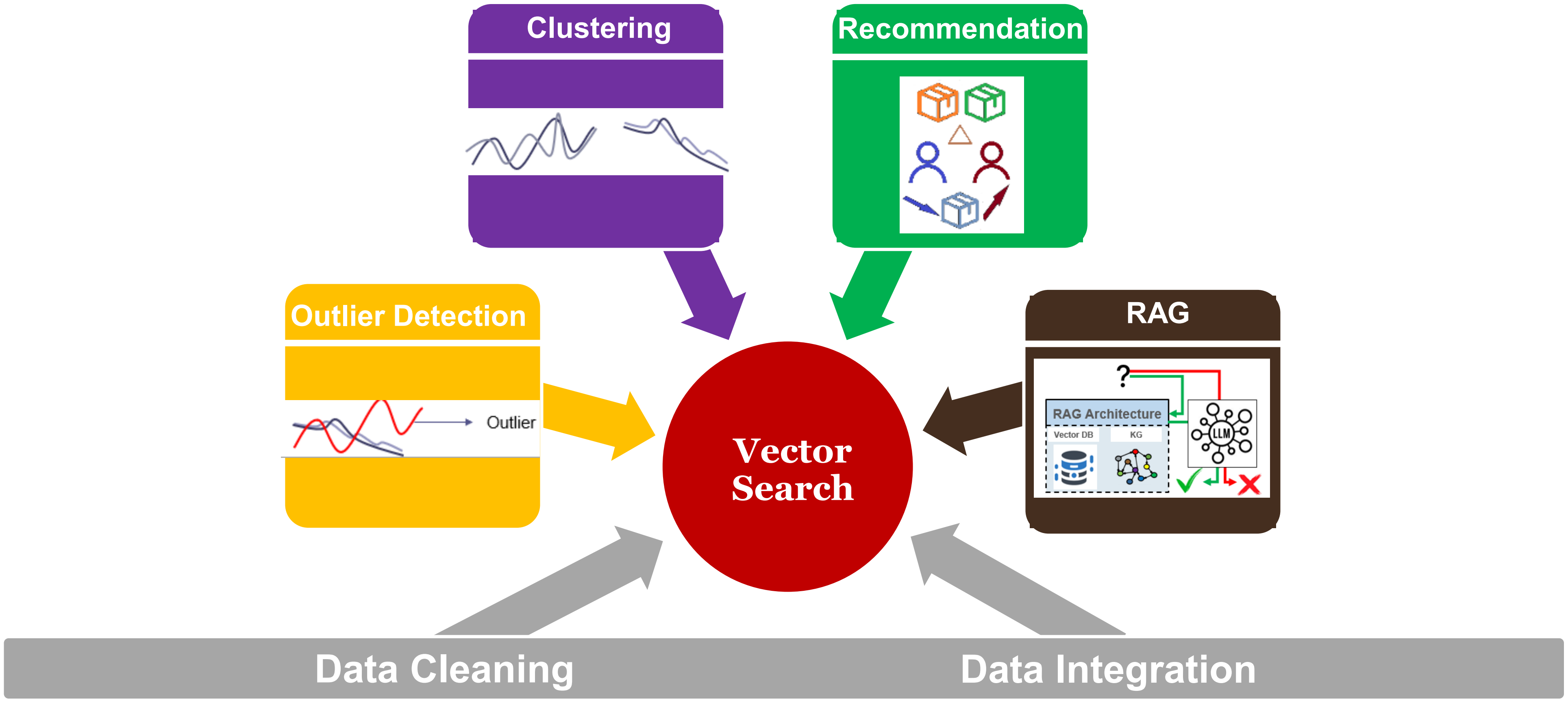
Vector Similarity Search at the Core of Data Science



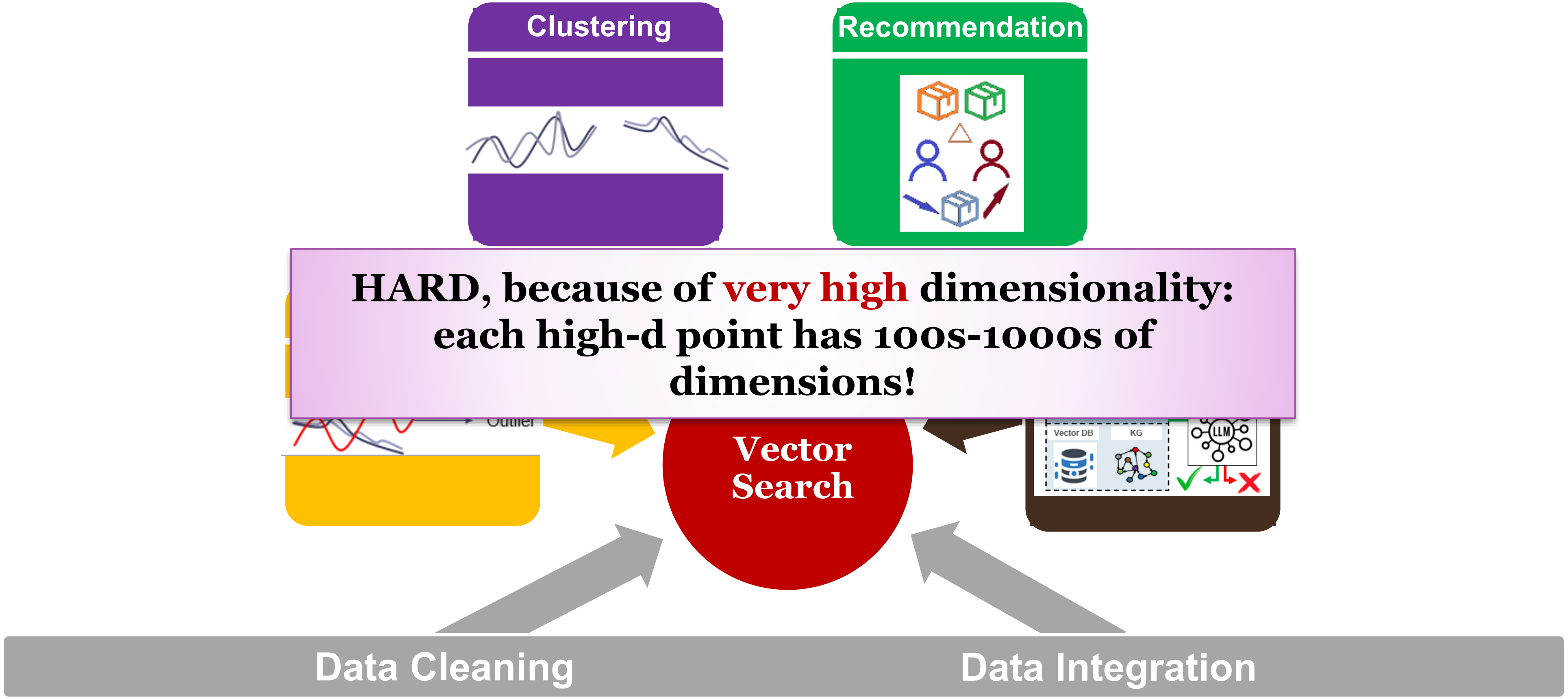
Vector Similarity Search at the Core of Data Science



Vector Similarity Search at the Core of Data Science



Vector Similarity Search at the Core of Data Science



HARD, because of very high dimensionality: each high-d point has 100s-1000s of dimensions!

Vector Similarity Search at the Core of Data Science

HARD, because of very high dimensionality: each high-d point has 100s-1000s of dimensions!

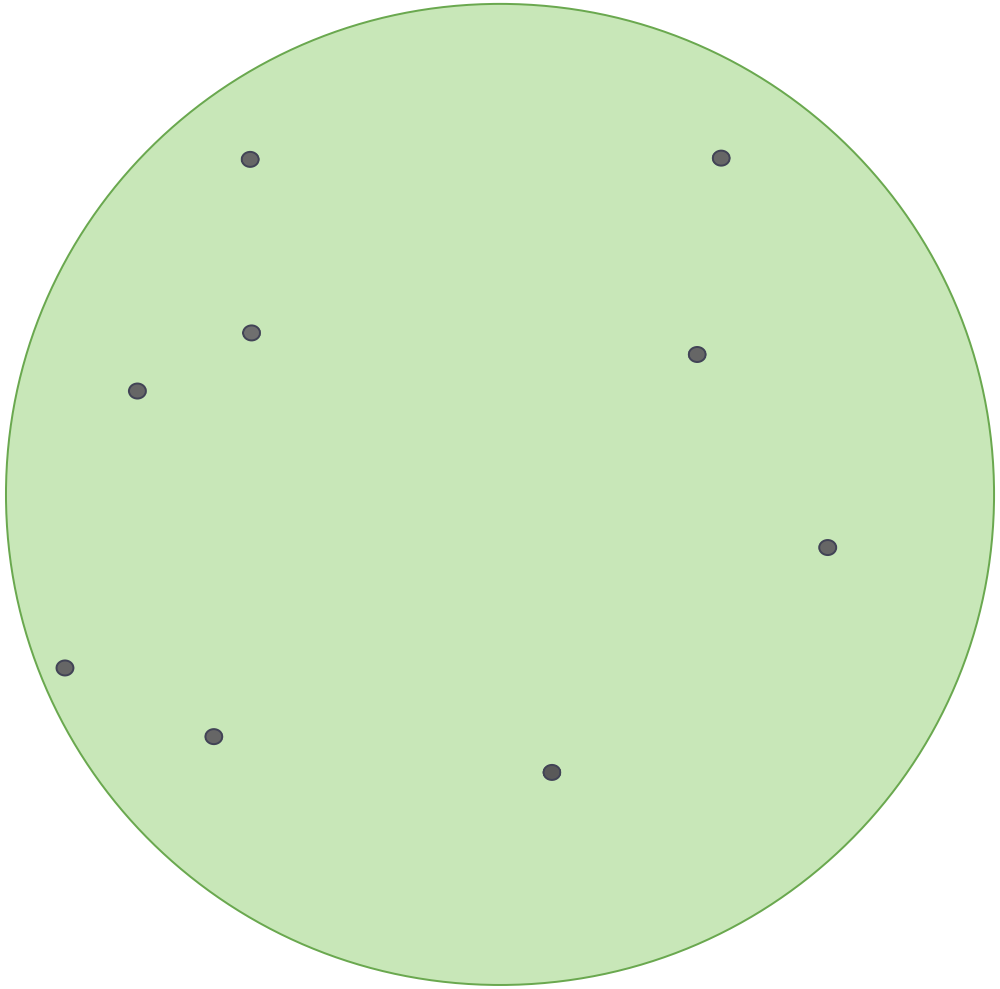
even HARDER, because of very large size: millions to billions of high-d points (multi-TBs)!

Data Cleaning | Data Integration

Nearest Neighbor (NN) Queries

Publications

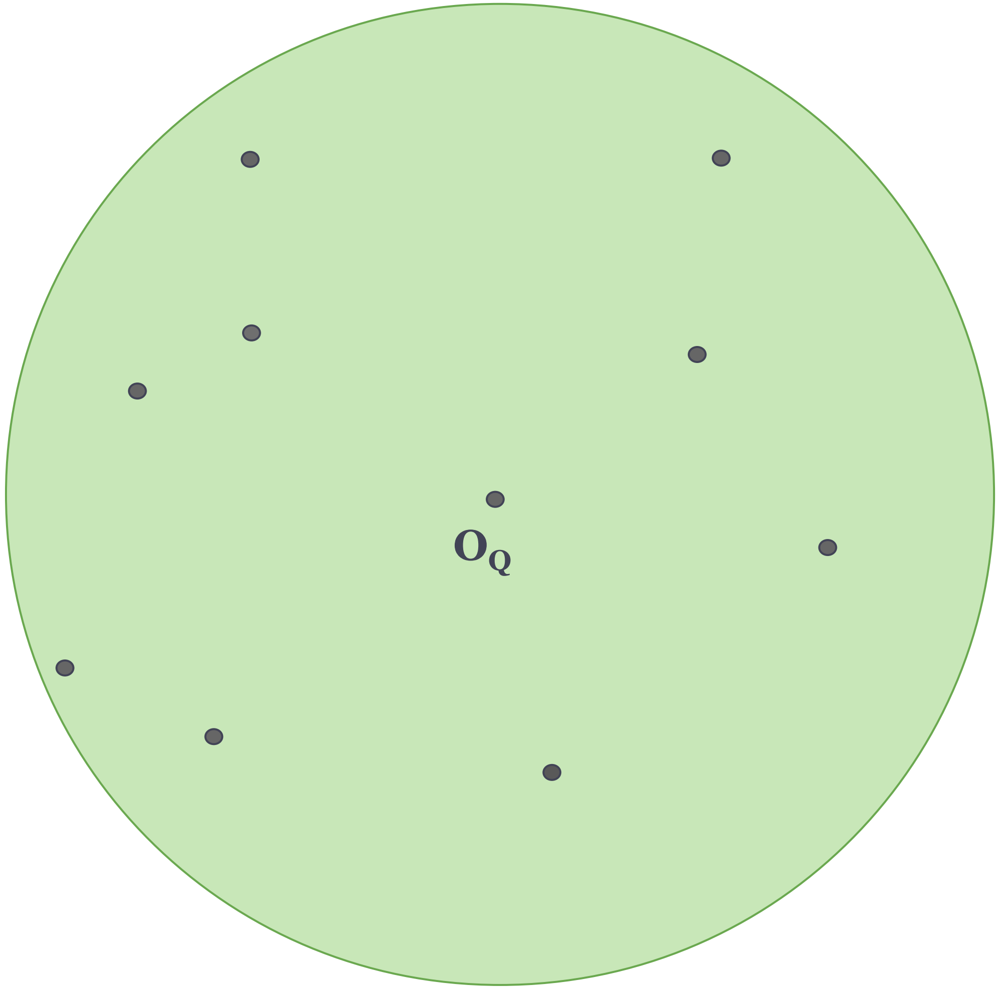
Echihabi et al. PVLDB'19



Nearest Neighbor (NN) Queries

Publications

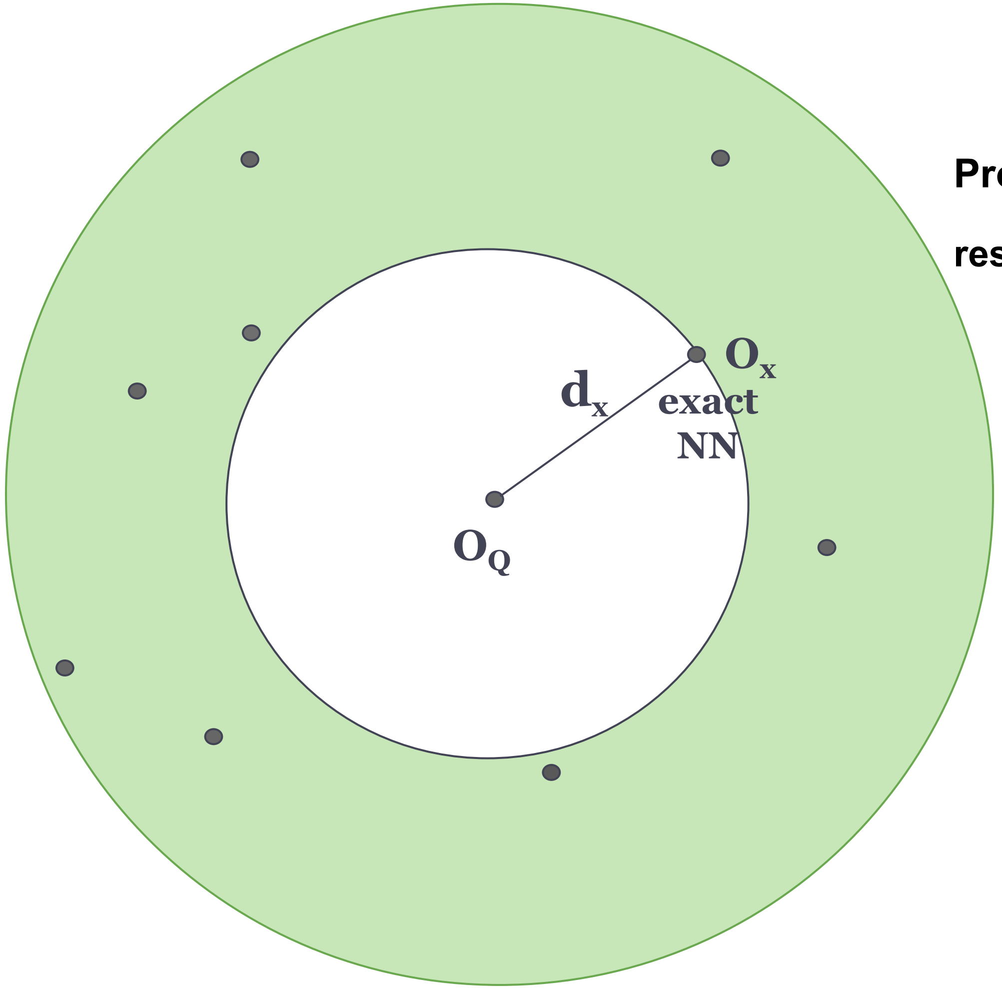
Echihabi et al. PVLDB'19



Nearest Neighbor (NN) Queries

Publications

Echihabi et al. PVLDB'19



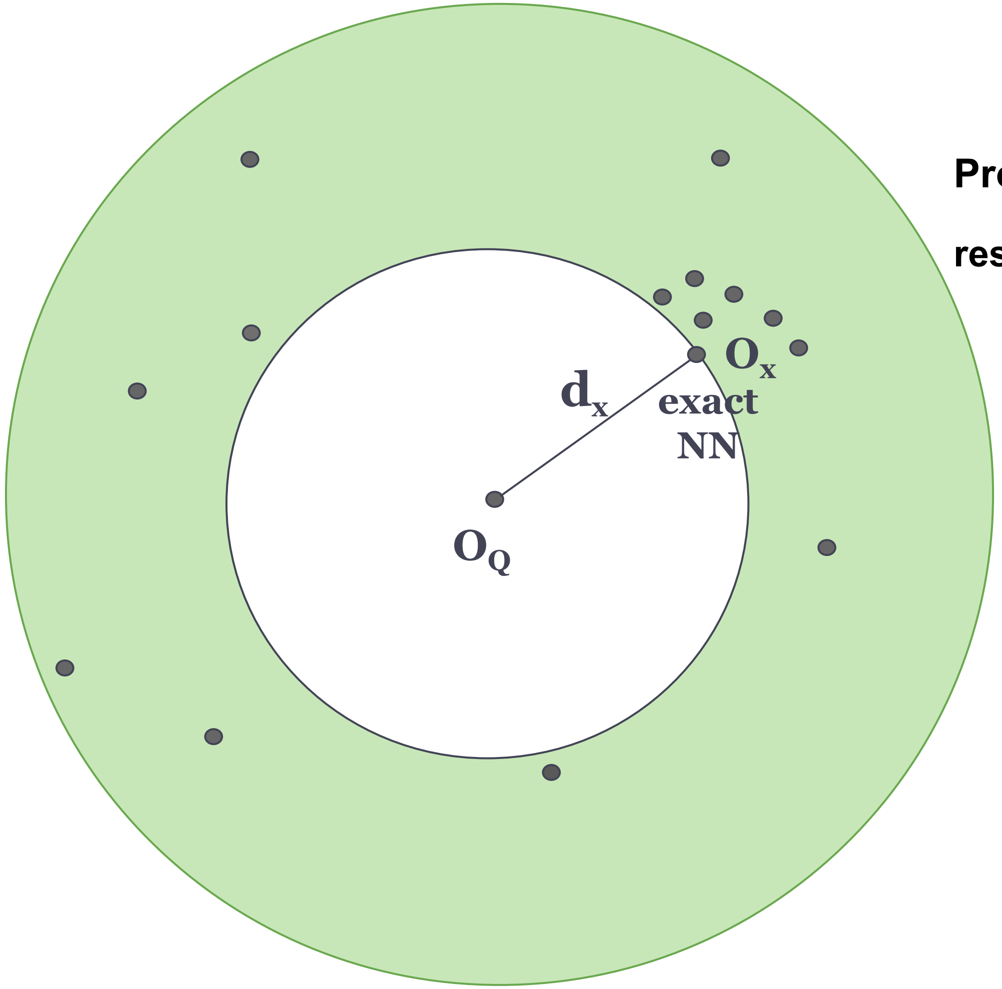
Prob($d_x = \min\{d_i\}$) = 1

result is exact NN

Nearest Neighbor (NN) Queries

Publications

Echihabi et al. PVLDB'19



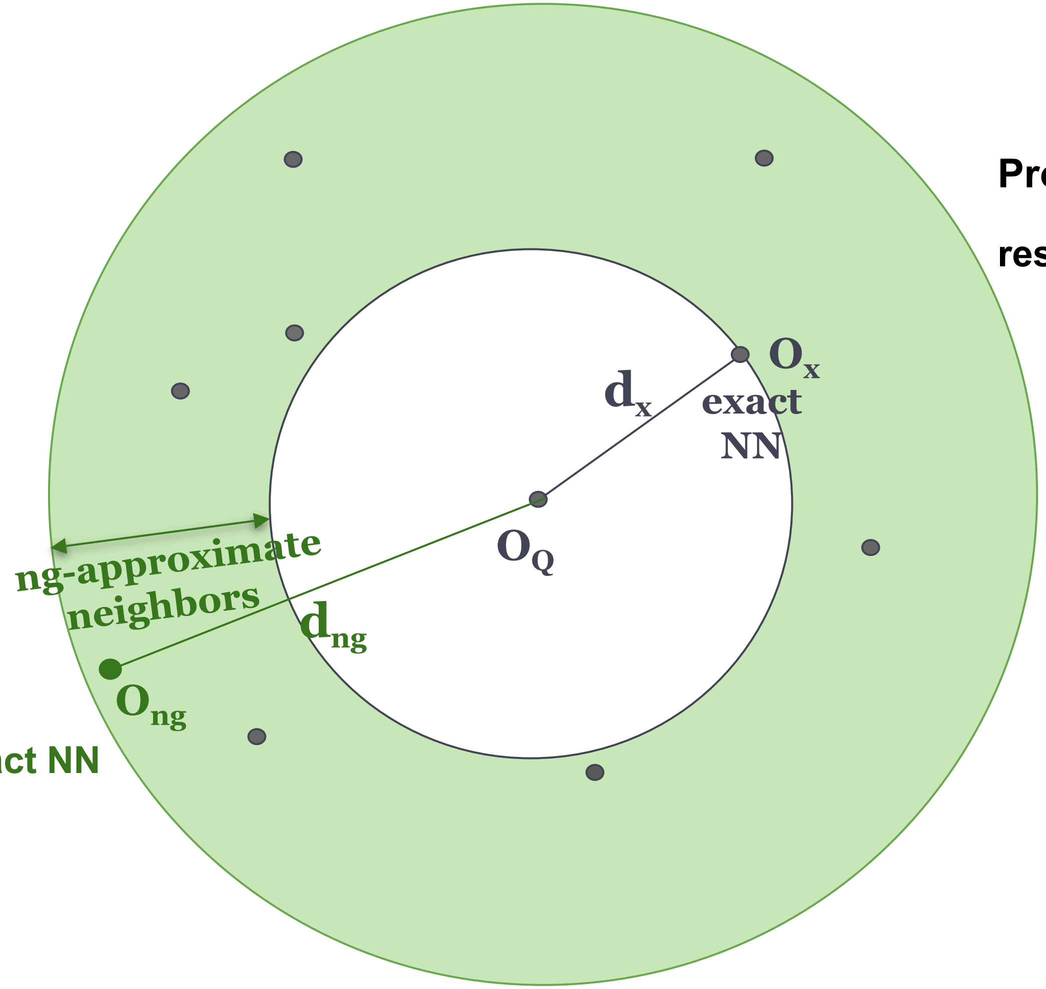
Prob($d_x = \min\{d_i\}$) = 1

result is exact NN

Nearest Neighbor (NN) Queries

Publications

Echihabi et al. PVLDB‘19



Prob($d_x = \min\{d_i\}$) = 1

result is exact NN

Prob(d_{ng} <>= ?) = ?

result within ? of exact NN

Nearest Neighbor (NN) Queries

Publications

Echihabi et al. PVLDB'19

Prob(d_ε <= d_x (1+ε)) = 1

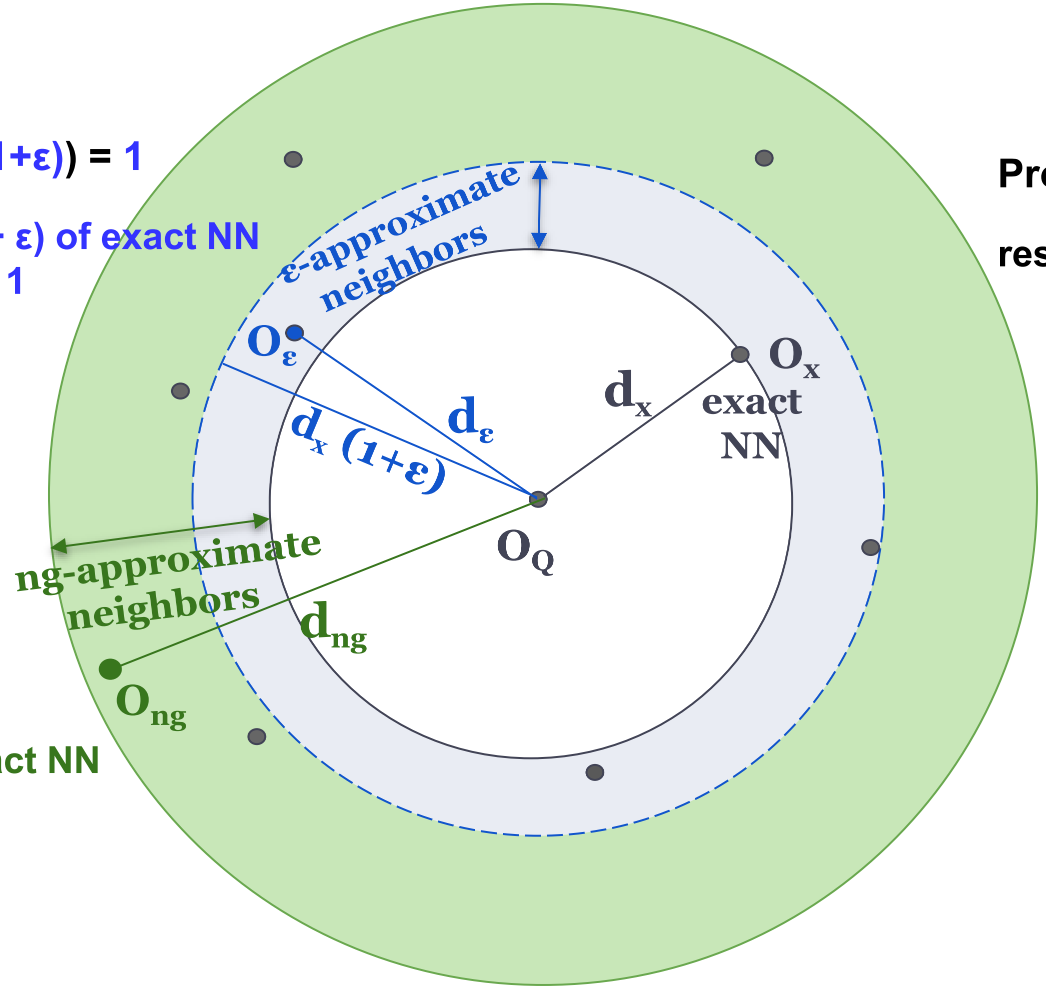
result within (1+ ε) of exact NN
with probability 1

Prob($d_x = \min\{d_i\}$) = 1

result is exact NN

Prob(d_{ng} <>= ?) = ?

result within ? of exact NN



Nearest Neighbor (NN) Queries

Publications

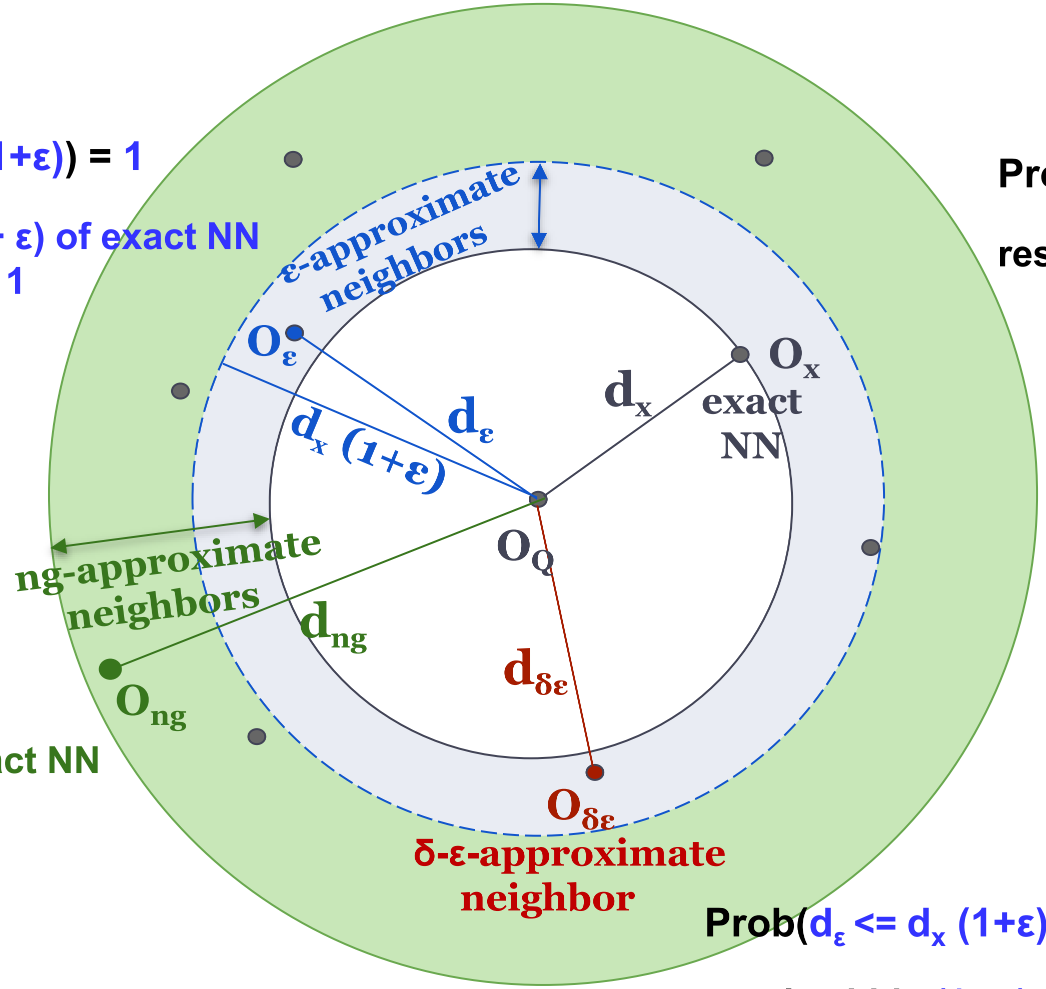
Echihabi et al. PVLDB‘19

Prob(d_ε <= d_x (1+ε)) = 1

result within (1+ ε) of exact NN
with probability 1

Prob(d_x = min{d_i}) = 1

result is exact NN



Prob(d_{ng} <>= ?) = ?

result within ? of exact NN

Prob(d_ε <= d_x (1+ε)) >= δ

result within (1+ ε) of exact NN
with probability at least δ

Vector Search

Main Contributions

The Past (Until ~ 2018)

- Large body of work (> 50 years)

The Past (Until ~ 2018)

- Large body of work (> 50 years)
- Exact search:
  - Different communities working in isolation
  - Data series vs generic high-d vectors
  - Typically scan (e.g. VA-file) or tree-based (e.g., KD-Tree, RTree)
  - High-d vector techniques scale to a few GBs over 10s of dimensions
  - Data series techniques scale to 100s GBs over 1000s of dimensions

Publications

- Salton-Am.Doc.'63
- Bentley-CACM'75
- Guttman-SIGMOD'84
- Samet-MorganK'06
- Blott-VLDB'08

The Past (Until ~ 2018)

- Large body of work (> 50 years)
- Exact search:
  - Different communities working in isolation
  - Data series vs generic high-d vectors
  - Typically scan (e.g. VA-file) or tree-based (e.g., KD-Tree, RTree)
  - High-d vector techniques scale to a few GBs over 10s of dimensions
  - Data series techniques scale to 100s GBs over 1000s of dimensions
- Approximate search:
  - with guarantees relatively slow
    - LSH family

Publications

- Salton-Am.Doc.'63
- Bentley-CACM'75
- Guttman-SIGMOD'84
- Samet-MorganK'06
- Blott-VLDB'08

Publications

- Kenneth-SICOMP'88
- Chin-Algorithmica'94
- Gionis-VLDB'99
- Jafari-Arxiv'21

The Past (Until ~ 2018)

- Large body of work (> 50 years)
- Exact search:
  - Different communities working in isolation
  - Data series vs generic high-d vectors
  - Typically scan (e.g. VA-file) or tree-based (e.g., KD-Tree, RTree)
  - High-d vector techniques scale to a few GBs over 10s of dimensions
  - Data series techniques scale to 100s GBs over 1000s of dimensions
- Approximate search:
  - with guarantees relatively slow
    - LSH family
  - without guarantees relatively efficient
    - Inverted indexes and graph-based techniques

Publications

- Salton-Am.Doc.'63
- Bentley-CACM'75
- Guttman-SIGMOD'84
- Samet-MorganK'06
- Blott-VLDB'08

Publications

- Kenneth-SICOMP'88
- Chin-Algorithmica'94
- Gionis-VLDB'99
- Jafari-Arxiv'21
- Jégou-TPAMI'11
- Malkov-IS'14

Meaningfulness of NN queries in high-d spaces

Publications

Beyer et al. ICDT'99

- Some studies have argued that NN search is not meaningful for a number of high dimensional datasets due to the concentration of distances.
  - However, these conclusions were based on over-restrictive assumptions

Meaningfulness of NN queries in high-d spaces

Publications

- Beyer et al. ICDT'99
- Aggarwal et al. ICDT'01
- He et al. ICML'12

- Some studies have argued that NN search is not meaningful for a number of high dimensional datasets due to the concentration of distances.
  - However, these conclusions were based on over-restrictive assumptions

- Other studies have shown that high-dimensional NN search is meaningful for:
  - non-i.i.d data
  - data with low intrinsic dimensionality
  - for a variety of real world datasets

The Present – Our Contributions

Publications

Echihabi-PVLDB'18

- Reconciled terminology/taxonomy across different communities

The Present – Our Contributions

Publications

Echihabi-PVLDB'18

- Reconciled terminology/taxonomy across different communities
- Exact search:
  - Data series techniques work well for generic high-d vectors
    - Evaluated on images, deep network embeddings.
    - Scaled to a few TBs of data over 1000s dimensions.

The Present – Our Contributions

Publications

- Echihabi-PVLDB‘18
- Gogolou-SIGMOD‘20
- Echihabi-VLDBJ‘23

- Reconciled terminology/taxonomy across different communities
- Exact search:
  - Data series techniques work well for generic high-d vectors
    - Evaluated on images, deep network embeddings.
    - Scaled to a few TBs of data over 1000s dimensions.
  - Progressive search
    - Proposed progressive search algorithms for interactive search.

The Present – Our Contributions

Publications

- Echihabi-PVLDB'18
- Gogolou-SIGMOD'20
- Echihabi-VLDBJ'23
- Echihabi-PVLDB'22

- Reconciled terminology/taxonomy across different communities
- Exact search:
  - Data series techniques work well for generic high-d vectors
    - Evaluated on images, deep network embeddings.
    - Scaled to a few TBs of data over 1000s dimensions.
  - Progressive search
    - Proposed progressive search algorithms for interactive search.
  - No technique was an overall winner
    - Proposed Hercules with state-of-the-art performance across all popular query workloads.

The Present – Our Contributions

- Reconciled terminology/taxonomy across different communities
- Exact search:
  - Data series techniques work well for generic high-d vectors
    - Evaluated on images, deep network embeddings.
    - Scaled to a few TBs of data over 1000s dimensions.
  - Progressive search
    - Proposed progressive search algorithms for interactive search.
  - No technique was an overall winner
    - Proposed Hercules with state-of-the-art performance across all popular query workloads.
  - An exact technique better suited for embeddings
    - Proposed a new state-of-the-art technique with logarithmic average-case/worst-case query time complexity.

Publications

- Echihabi-PVLDB‘18
- Gogolou-SIGMOD‘20
- Echihabi-VLDBJ‘23
- Echihabi-PVLDB‘22
- Abdenouri-Submission

The Present – Our Contributions

Publications

Echihabi-PVLDB'19

- Approximate search:
  - Provided techniques that answer all flavors of search and are alternatives to:
    - the LSH family for approximate search with guarantees based on trees.
    - graph-based and quantization-based methods for search without guarantees on disk.

The Present – Our Contributions

Publications

Echihabi-PVLDB'19

Azizi-PVLDB'23

- Approximate search:
  - Provided techniques that answer all flavors of search and are alternatives to
    - the LSH family for approximate search with guarantees based on trees.
    - graph-based and quantization-based methods for search without guarantees on disk.
  - Proposed Elpis to address the indexing scalability of graph-based techniques
    - builds the index 3x-8x faster than competitors, using 40% less memory.
    - achieves a high recall of 0.99, up to 2x faster than the state-of-the-art methods.

The Present – Our Contributions

- Approximate search:
  - Provided techniques that answer all flavors of search and are alternatives to
    - the LSH family for approximate search with guarantees based on trees.
    - graph-based and quantization-based methods for search without guarantees on disk.
  - Proposed Elpis to address the indexing scalability of graph-based techniques
    - builds the index 3x-8x faster than competitors, using 40% less memory.
    - achieves a high recall of 0.99, up to 2x faster than the state-of-the-art methods.
  - Conducted an experimental evaluation of in-memory graph-based vector indexes
    - Incremental insertion and neighborhood diversification lead to better query performance.
    - Efficient seed selection can improve both indexing and search performance.

Publications

Echihabi-PVLDB'19

Azizi-PVLDB'23

Azizi-PACMMOD'25

The Present – Our Contributions

- Approximate search:
  - Provided techniques that answer all flavors of search and are alternatives to
    - the LSH family for approximate search with guarantees based on trees.
    - graph-based and quantization-based methods for search without guarantees on disk.
  - Proposed Elpis to address the indexing scalability of graph-based techniques
    - builds the index 3x-8x faster than competitors, using 40% less memory.
    - achieves a high recall of 0.99, up to 2x faster than the state-of-the-art methods.
  - Conducted an experimental evaluation of in-memory graph-based vector indexes
    - Incremental insertion and neighborhood diversification lead to better query performance.
    - Efficient seed selection can improve both indexing and search performance.
  - Proposed RWalks, an index-agnostic graph-based filtered vector search method
    - Efficiently supports both filtered and unfiltered vector search.
    - Can perform filtered search up to 2x faster than the second-best competitor (ACORN), while building the index 76x faster and answering unfiltered search 13x faster.

Publications

Echihabi-PVLDB‘19

Azizi-PVLDB‘23

Azizi-PACMMOD‘25

AitAomar-PACMMOD‘25

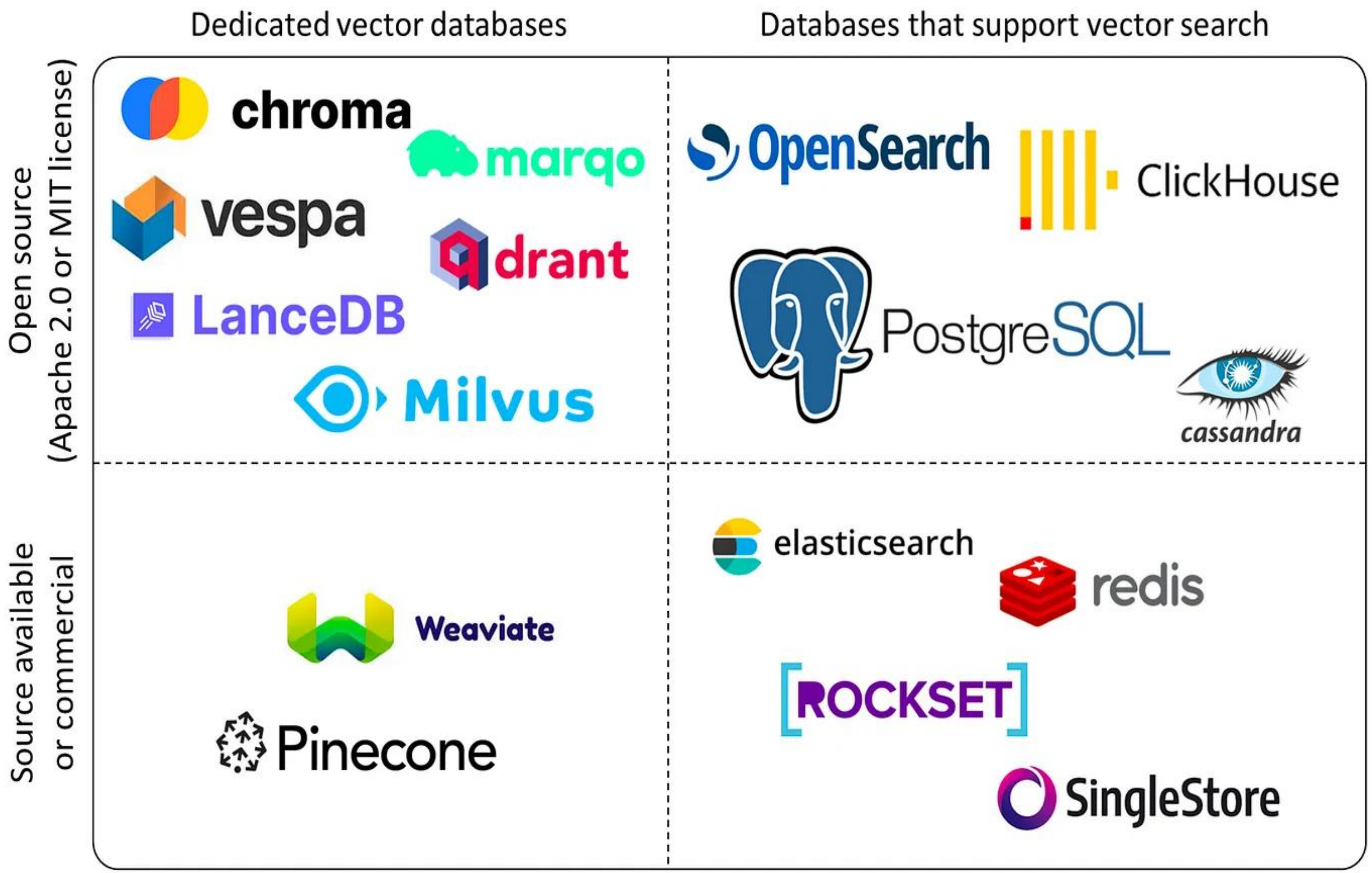
Publications

Azizi-PACMMOD‘25

Vector Search

Experimental Evaluation of Graph-Based Methods

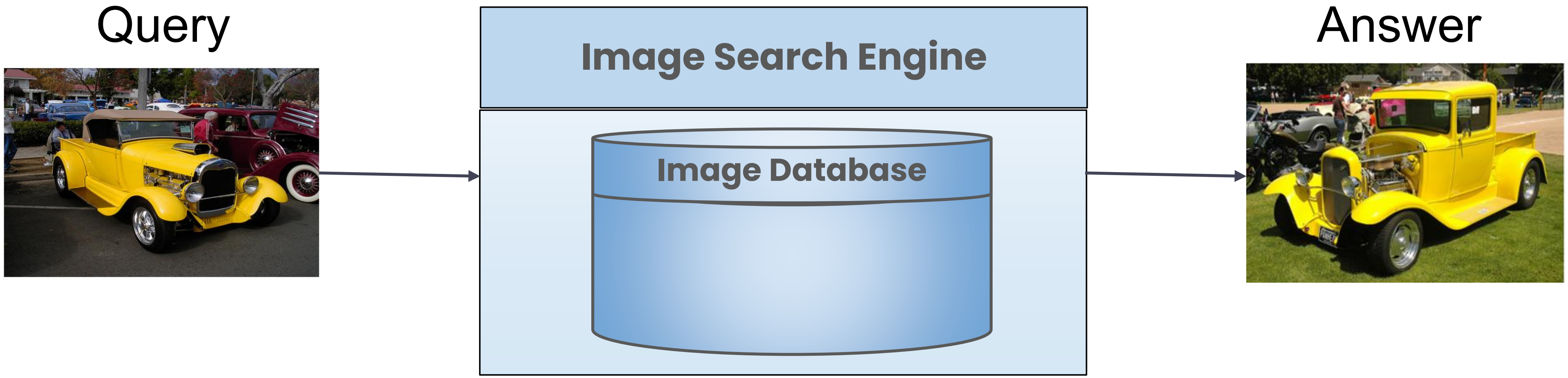
Why Graph-Based Search?



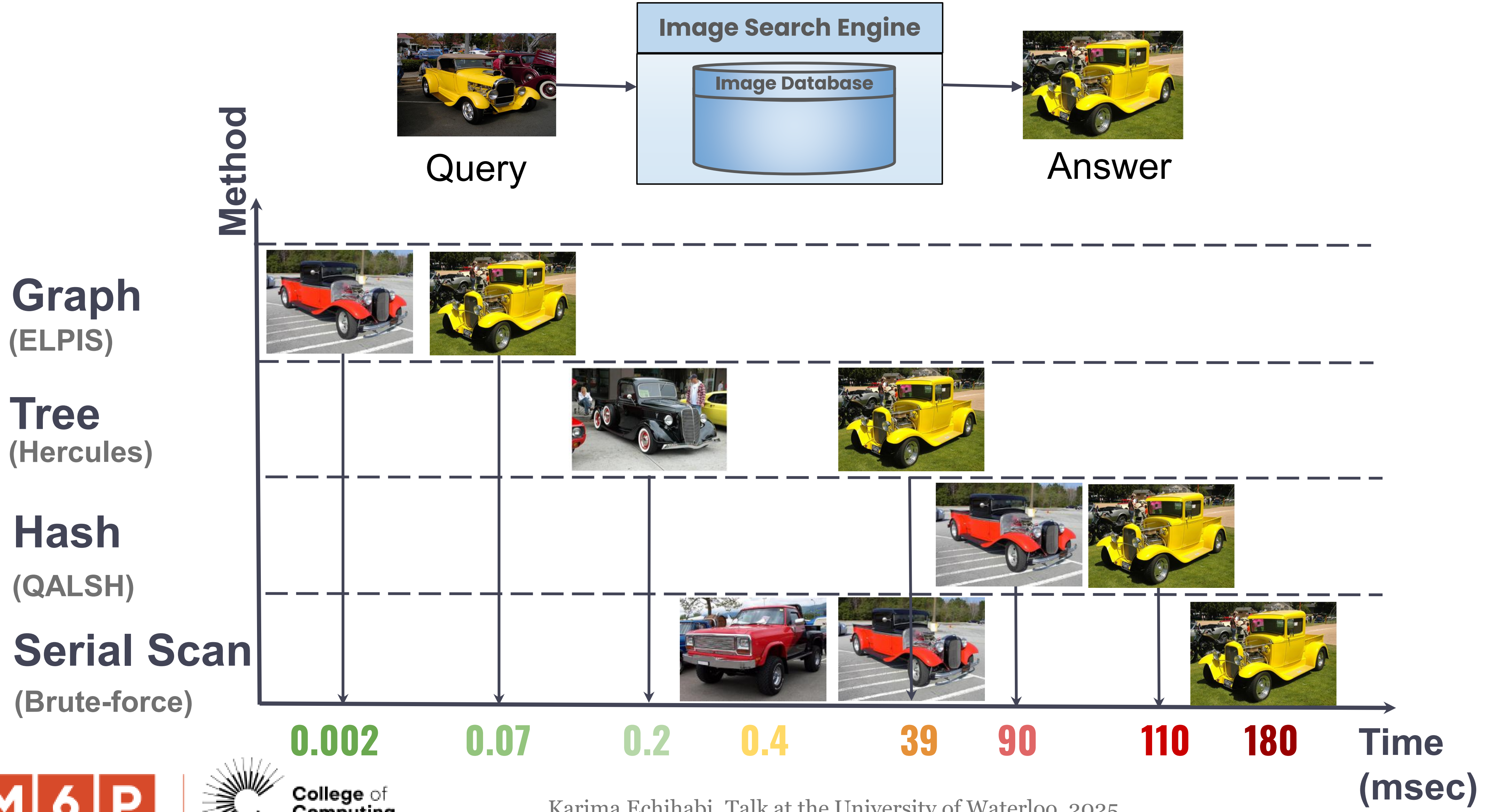
Why Graph-Based Search?

Pinecone	Proprietary composite index
milvus / zilliz	Flat, Annoy, IVF, HNSW/RHNSW (Flat/PQ), DiskANN
Weaviate	Customized HNSW, HNSW (PQ), DiskANN (in progress...)
Qdrant	Customized HNSW
chroma	HNSW
LanceDB	IVF (PQ), DiskANN (in progress...)
vespa	HNSW + BM25 hybrid
Vald	NGT
elasticsearch	Flat (brute force), HNSW
redis	Flat (brute force), HNSW
pgvector	IVF (Flat), IVF (PQ) in progress...

Why Graph-Based Search?



Why Graph-Based Search?



Limitations of Existing Studies

- Lack of comprehensive taxonomy
- Limited insights into graph construction’s impact on search performance.
- Limited scalability study (up to 1M vectors)

Contributions

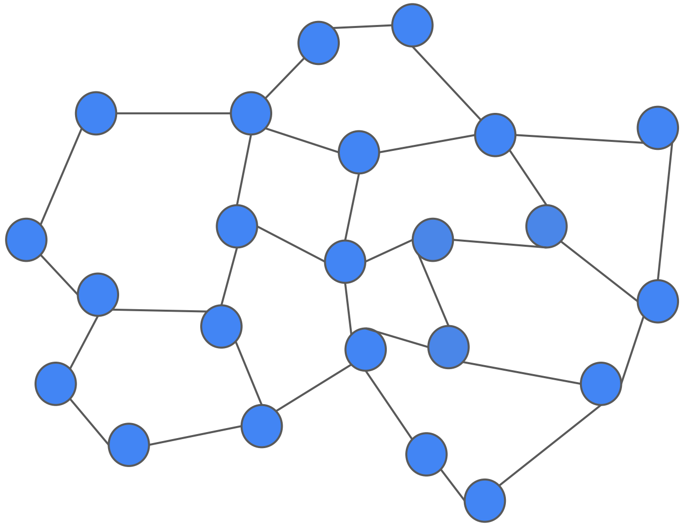
- New taxonomy of five design paradigms
- New insights on the impact of key design choices on performance
- Exhaustive experimental evaluation
- Recommendations
- Promising research directions

Primer - Proximity Graphs

Publications

Mitchell-Handbook‘17

Proximity graphs are geometric graphs $G(V,E)$ in which two vertices p, q are connected by an edge (p, q) if and only if a certain exclusion region for p, q contains no points from the vertex set.

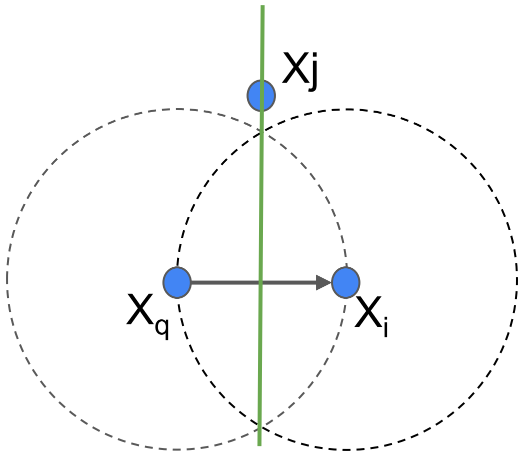


Primer – Example Proximity Graphs

Publications

Mitchell-Handbook'17

Relative Neighborhood Graph (RNG)
Lune (Xq,Xi) is empty

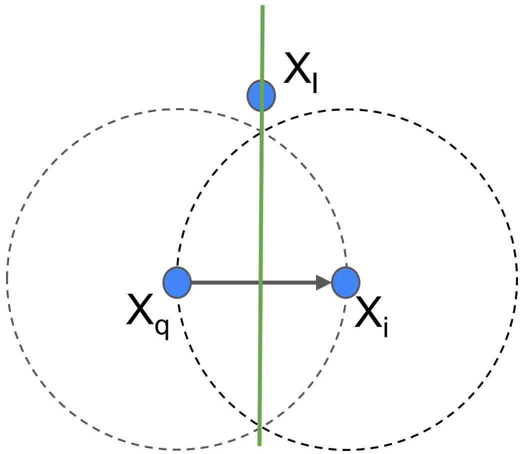


Primer – Example Proximity Graphs

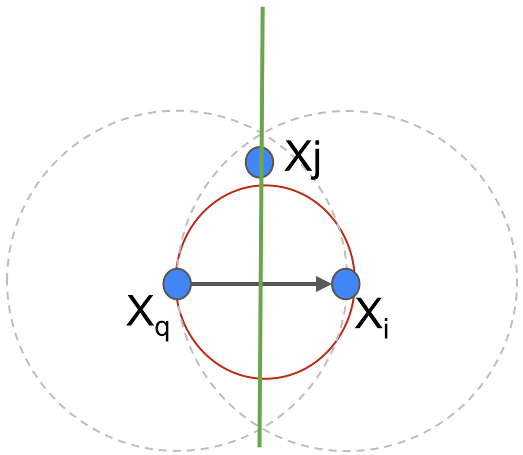
Publications

Mitchell-Handbook‘17

Relative Neighborhood Graph (RNG)
Lune (Xq,Xi) is empty



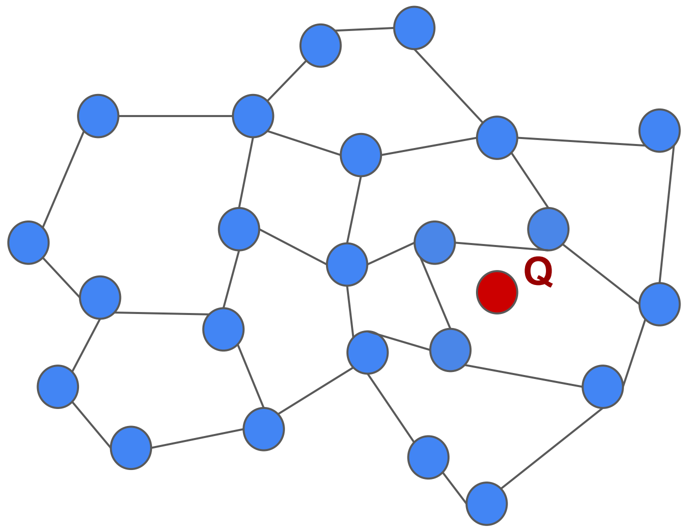
Gabriel Graph (GG) that is not RNG
DiameterSphere (Xq,Xi) is empty



Primer - Beam Search

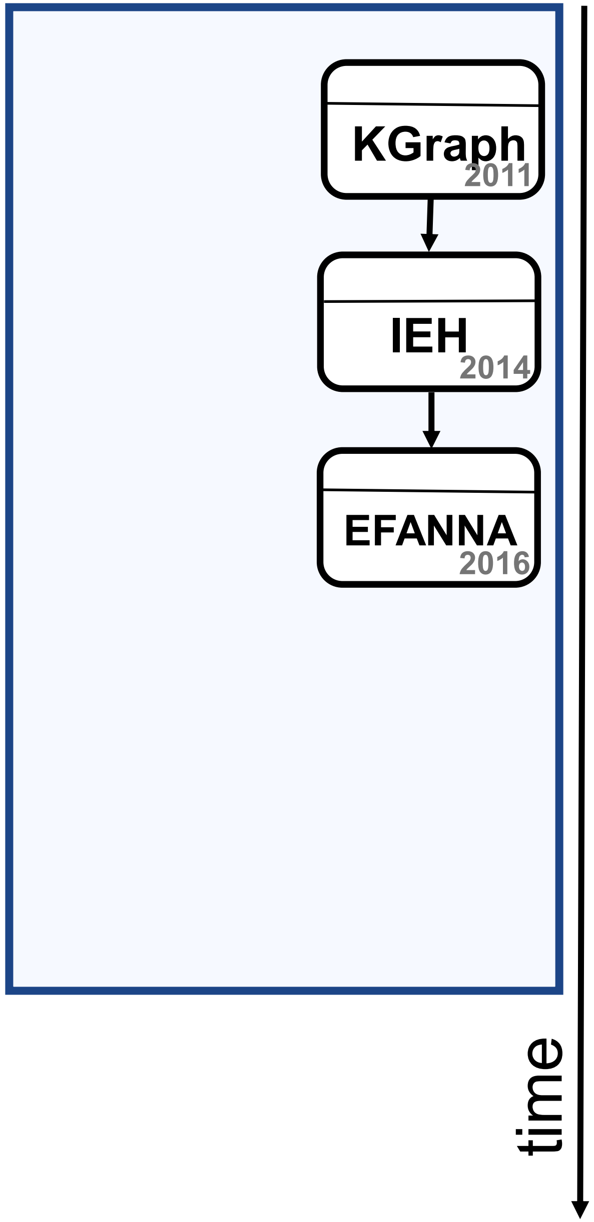
Beam search is a heuristic search algorithm that explores a graph by expanding the most optimistic node in a limited set of size L

BeamSearch (G, Q, entry_node, K, L)



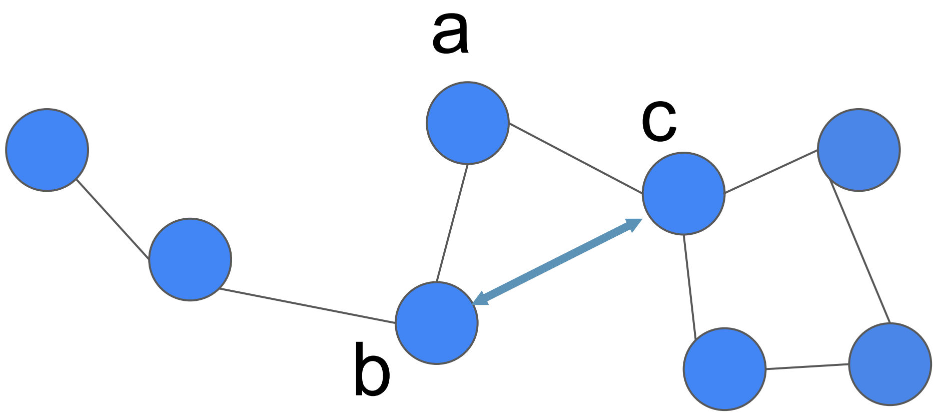
Proposed Taxonomy

Neighborhood Propagation



Neighborhood Propagation

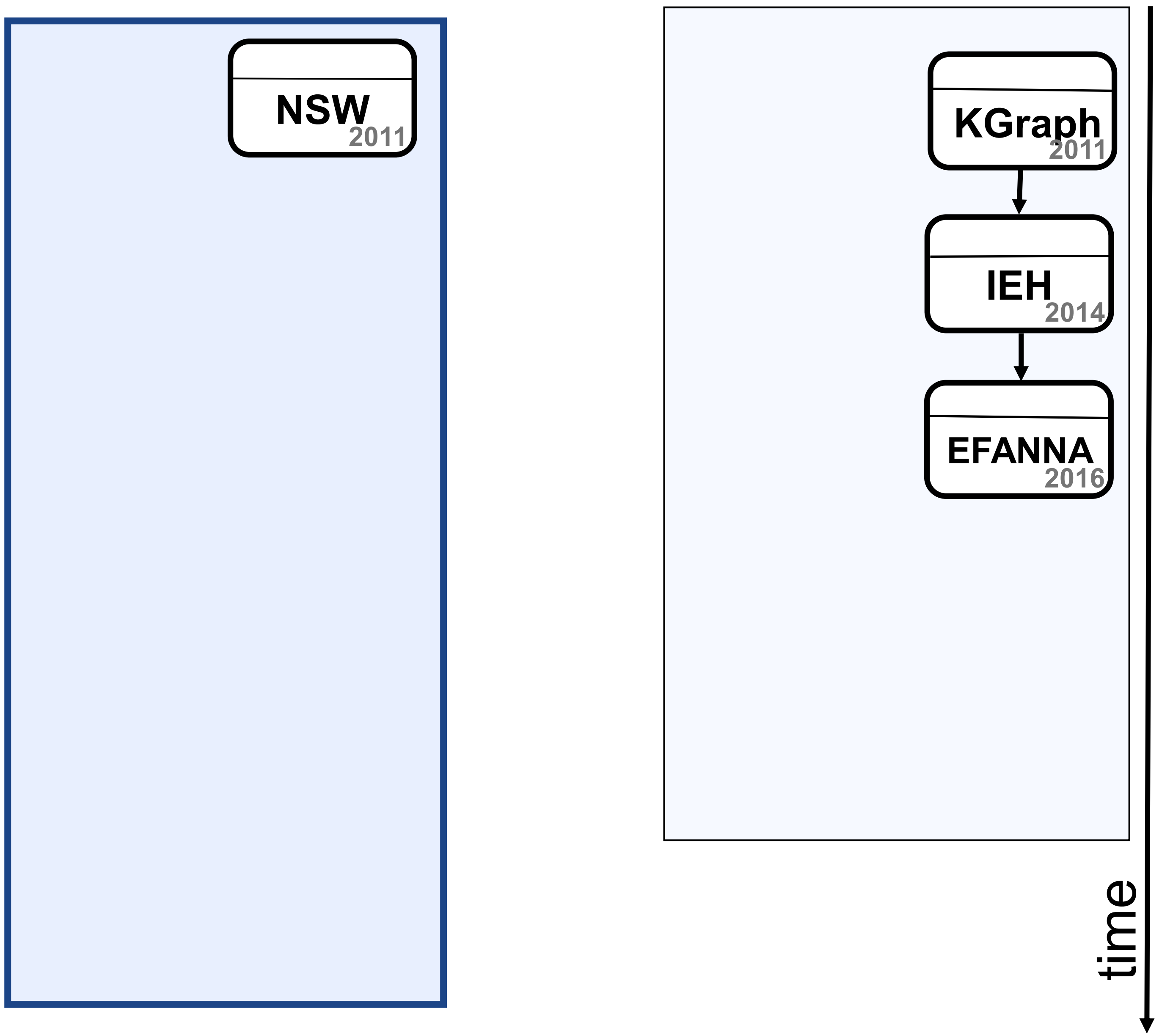
- Neighborhood propagation through NNDescent



“Hi a, your neighbors can become my neighbors too”

Proposed Taxonomy

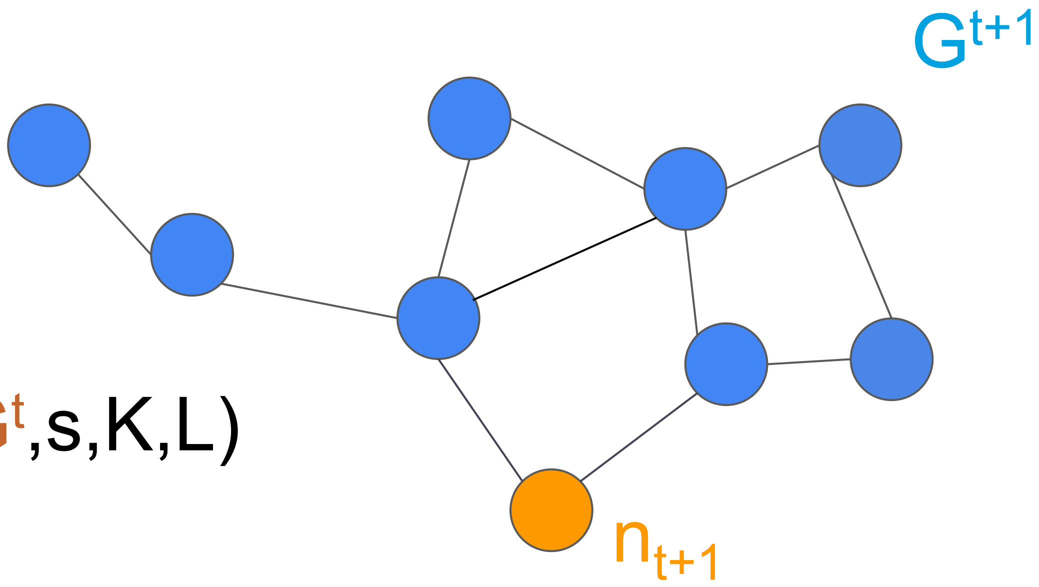
Neighborhood Propagation
Incremental Insertion



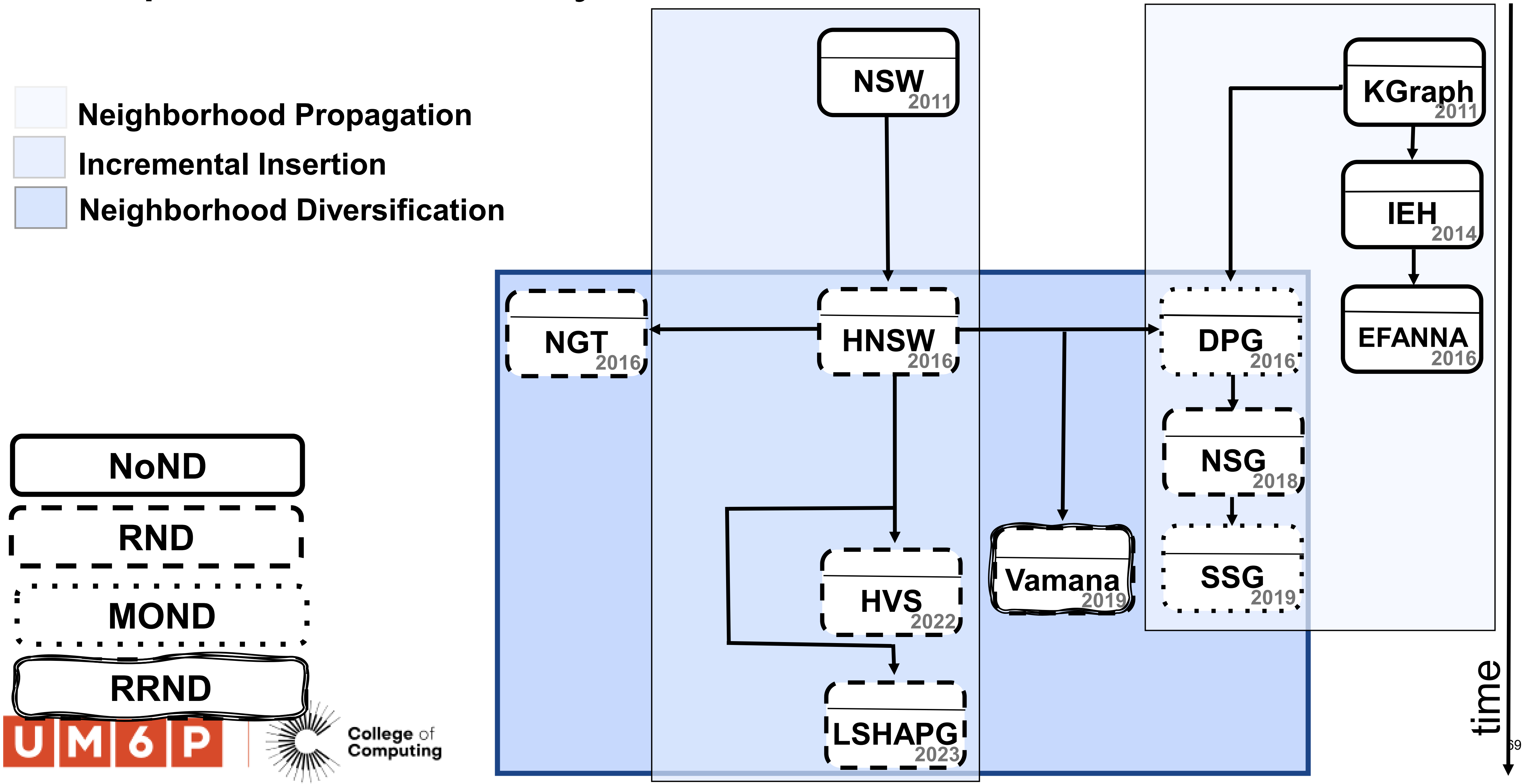
Incremental Insertion

- Insert node incrementally into the graph

BeamSearch (n_{t+1},G^t,s,K,L)



Proposed Taxonomy



UM6P College of Computing

Relative Neighborhood Diversification

For a node $\mathbf{X_q}$ and a list of candidate neighbors $\mathbf{C_q}$, the node $\mathbf{X_j}$, which belongs to $\mathbf{C_q}$, is selected into the set of $\mathbf{X_q}$'s neighbors $\mathbf{R_q}$ if and only if the following condition holds:

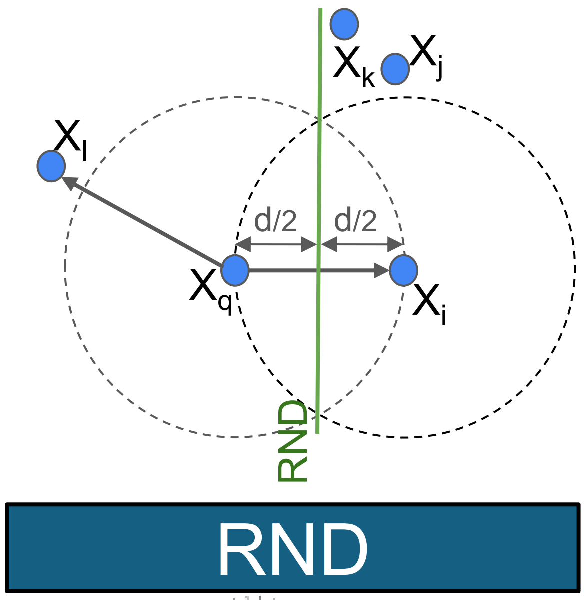
$$\forall \mathbf{X_i} \in \mathbf{R_q},\ \mathrm{dist}(\mathbf{X_q},\mathbf{X_j}) < \mathrm{dist}(\mathbf{X_i},\mathbf{X_j}) \quad \text{(eq1)}$$

Where:

- $\mathbf{X_q}$ is the query node.
- $\mathbf{X_j}$ is a candidate neighbor being considered for inclusion in $\mathbf{R_q}$ and is part of $\mathbf{C_q}$.
- $\mathbf{X_i}$ are nodes already in the set $\mathbf{R_q}$.
- $\mathrm{dist}(\mathbf{X_a},\mathbf{X_b})$ represents the Euclidean distance between nodes $\mathbf{X_a}$ and $\mathbf{X_b}$ in the d-dimensional space.

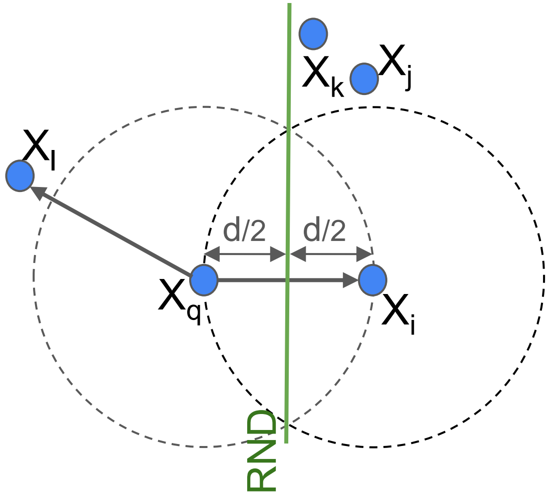
Neighborhood Diversification Today

1. Relative ND (RND) → approximate RNG as |Cq| << |V|

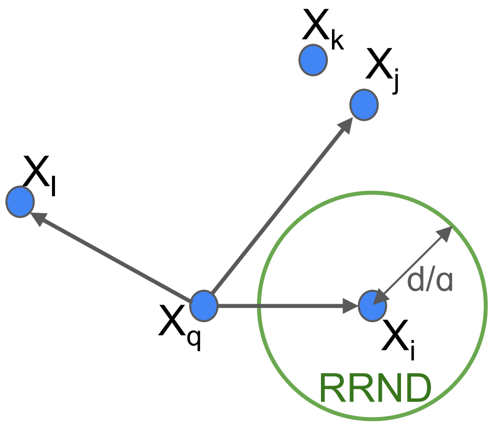


Neighborhood Diversification Today

1. Relative ND (RND)
2. Relaxed RND (RRND): (eq_1) => $\forall \mathbf{X_j} \in \mathbf{R_q}$, dist($\mathbf{X_q}$,$\mathbf{X_j}$) < α.dist($\mathbf{X_i}$,$\mathbf{X_j}$) for α > 1



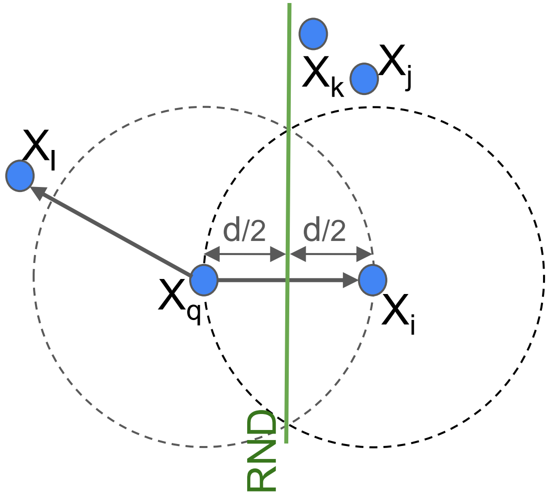
RND



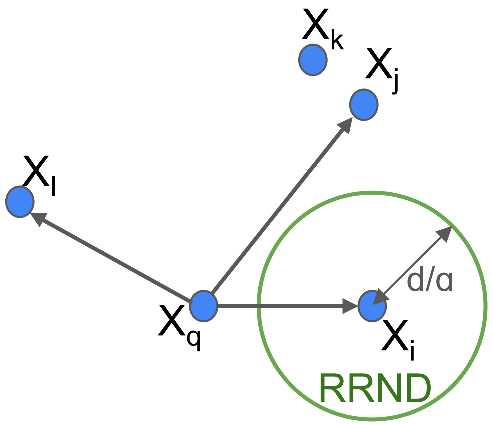
RRND

Neighborhood Diversification Today

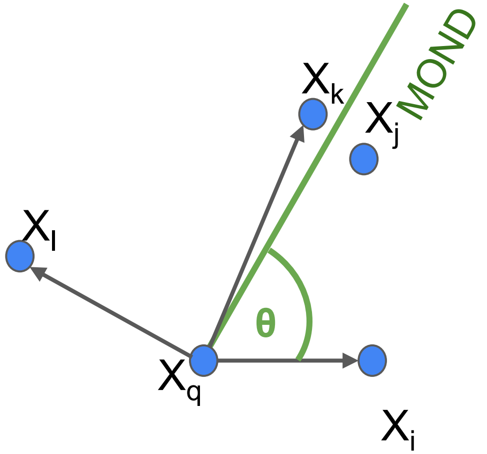
1. Relative ND (RND)
2. Relaxed RND (RRND): (eq_1) => $\forall \mathbf{X_j} \in \mathbf{R_q}$, $dist(\mathbf{X_q},\mathbf{X_j}) < \alpha . dist(\mathbf{X_i},\mathbf{X_j})$ for $\alpha > 1$
3. Maximum-Oriented ND (MOND): (eq_1) => $\forall \mathbf{X_j} \in \mathbf{R_q}$, $\cos(\angle \mathbf{X_i X_q X_j}) < \cos(\theta)$



RND

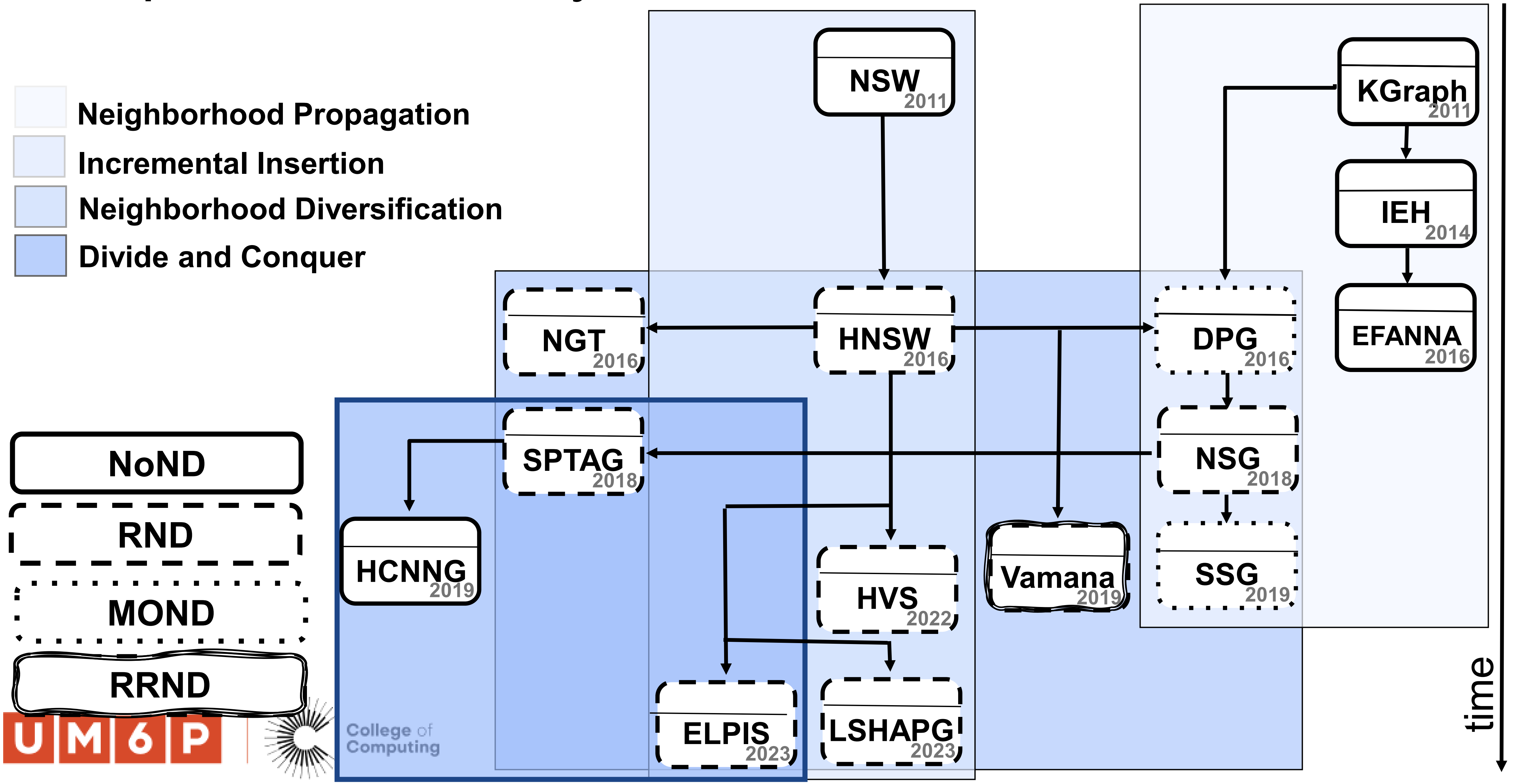


RRND



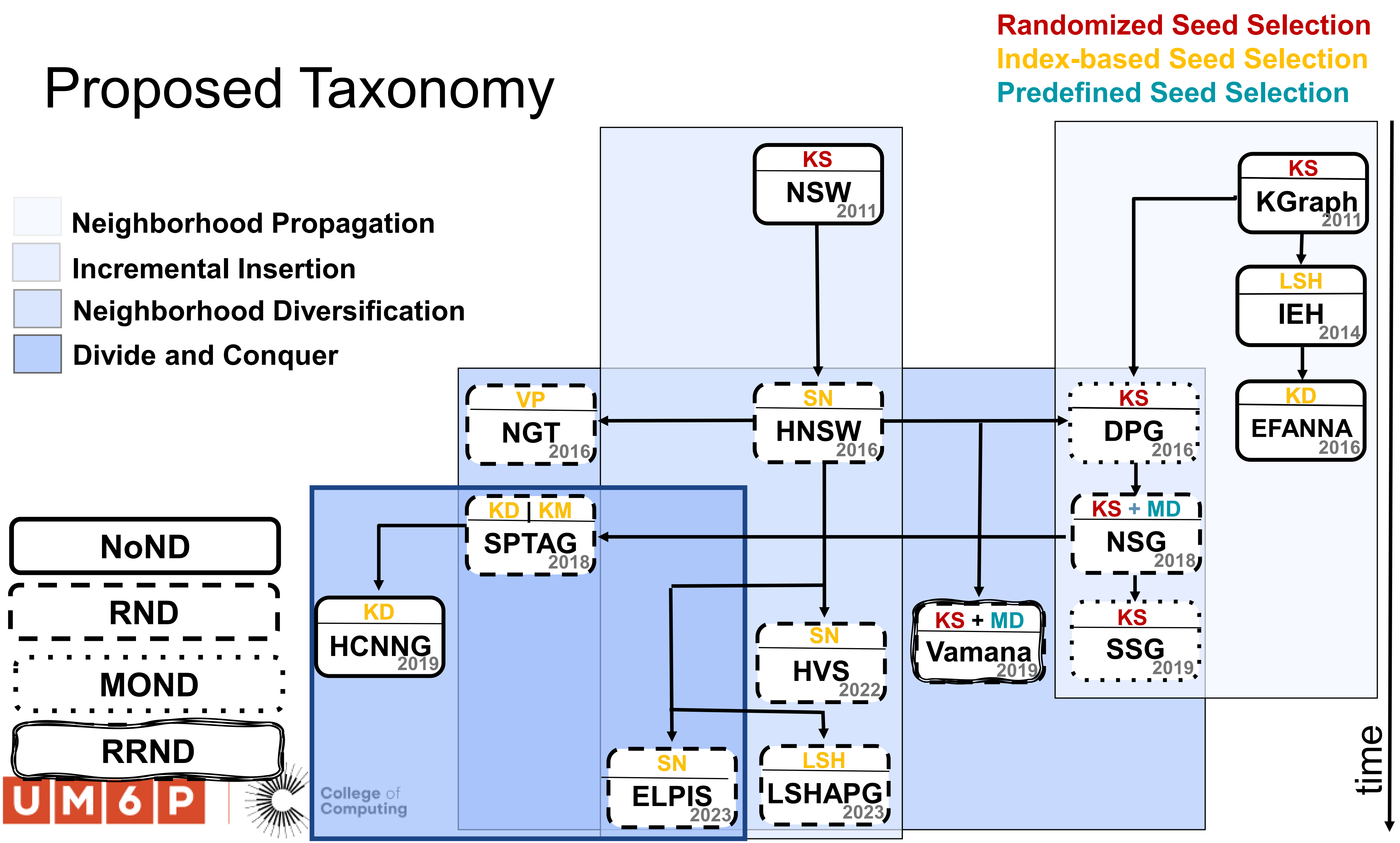
MOND

Proposed Taxonomy



College of Computing

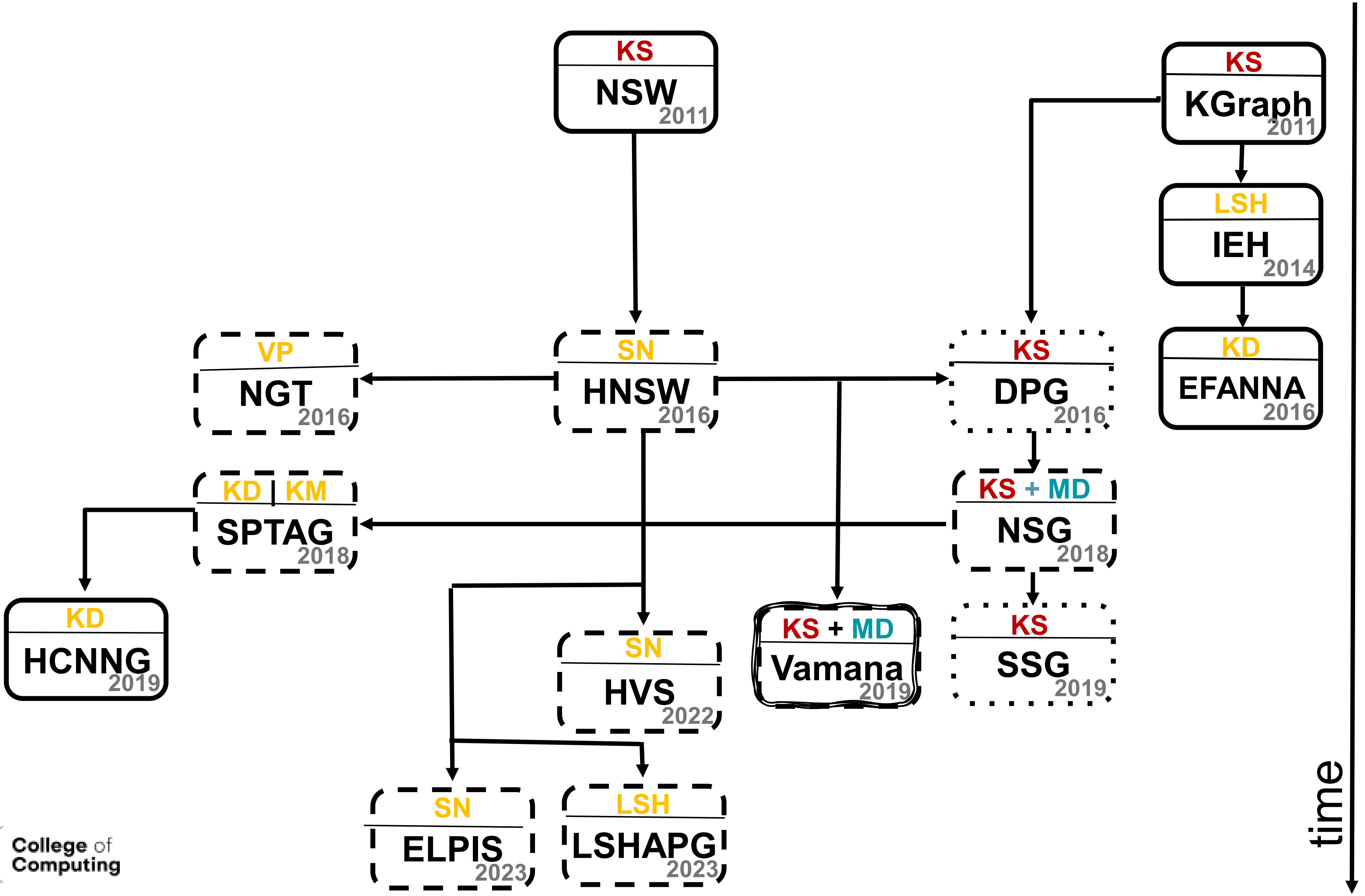
Proposed Taxonomy



UM6P College of Computing

Seed Selection

Randomized Seed Selection
Index-based Seed Selection
Predefined Seed Selection



Experimental Evaluation – Datasets

- **Sift1B**: 1 billion vectors of 128 dimensions representing the Sift image feature descriptors.
- **Deep1B**: 1 billion vectors of 96 dimensions extracted from the last layers of a convolutional neural network.
- **Sald**: 200 million neuroscience MRI data series of size 128 .
- **Seismic**: 100 million data series of size 256 representing earthquake recordings at seismic stations worldwide.
- **Gist**: 1 million vectors of 960 dimensions representing image descriptors that capture spatial structure and color layout.
- **ImageNet**: 1 million vectors of 256 dimensions generated from ImageNet using a ResNet50 model, followed by PCA for dimensionality reduction.
- **Text-to-Image**: 1 billion 200-dimensional image embeddings (from Se-ResNext-101) paired with 50 million text queries (from DSSM) for cross-modal retrieval tasks.
- **RandPow*i***: contains vectors of 256 dimensions generated randomly following power law distribution using power law exponent *i*.

Experimental Evaluation – Query Workloads

- Query sets include 100 vectors processed sequentially, not in batches, mimicking a real-world scenario where queries are unpredictable.
- Results with 1 million queries are extrapolated from 100 query sets.
- For Deep, Sift, GIST, and Text-to-Image, queries are randomly sampled from available query workloads.
- For SALD, ImageNet, and Seismic, For SALD, ImageNet, and Seismic, 100 queries are randomly selected from the datasets and excluded from the index-building phase.
- For hardness experiments, we use Deep query vectors of varying complexity, denoted as a percentage ranging from 1% to 10% (percentage indicating the gaussian noise added).

Experimental Evaluation – Hard Query Workloads

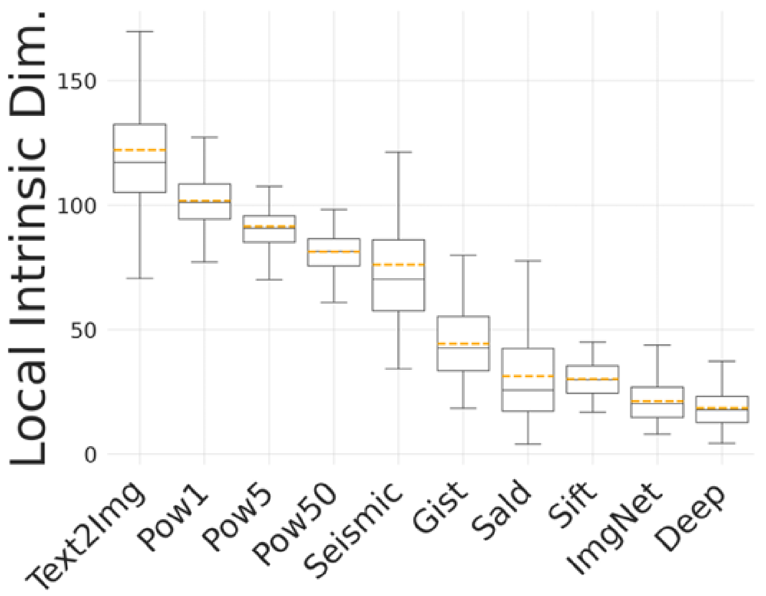
- Local Intrinsic Dimensionality (LID):

$$\mathrm{LID}(x) = -\left(\frac{1}{k}\sum_{i=1}^{k}\log\frac{\mathrm{dist}_i(x)}{\mathrm{dist}_k(x)}\right)^{-1}$$

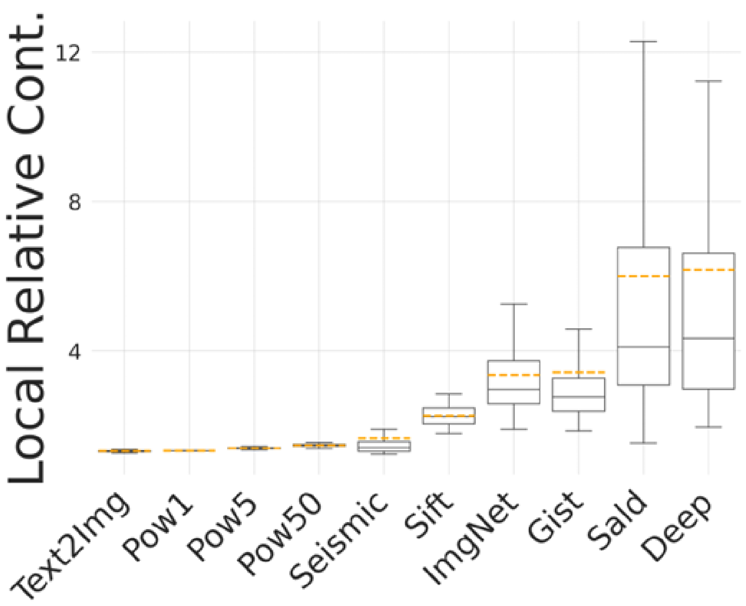
- Local Relative Contract:

$$\mathrm{LRC}(x) = \frac{\mathrm{dist}_{\mathrm{mean}}(x)}{\mathrm{dist}_k(x)}$$

Experimental Evaluation – Query Workloads



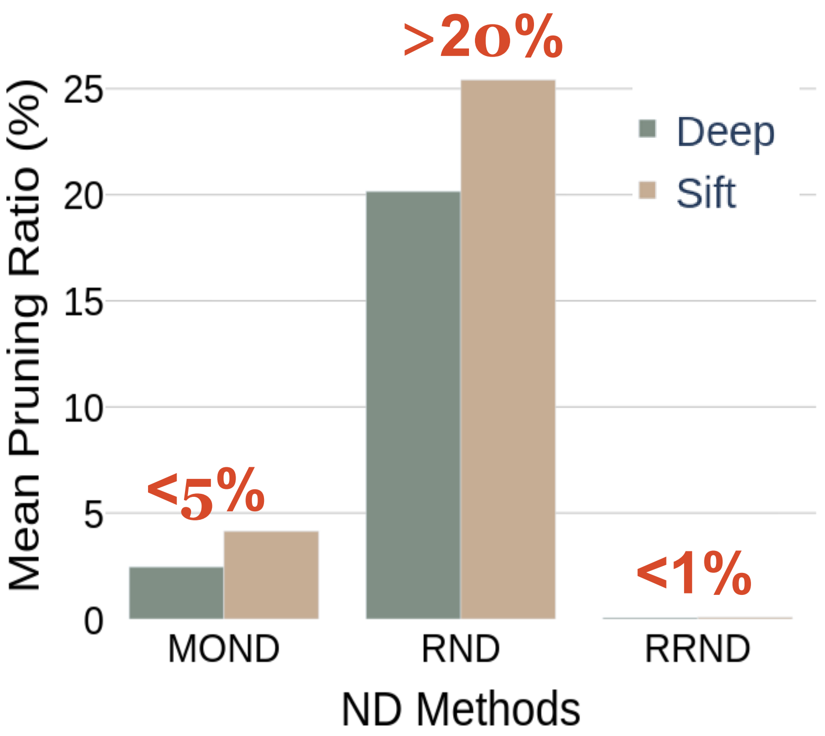
(a) Local Intrinsic Dimensionality (LID): low values indicate easy search

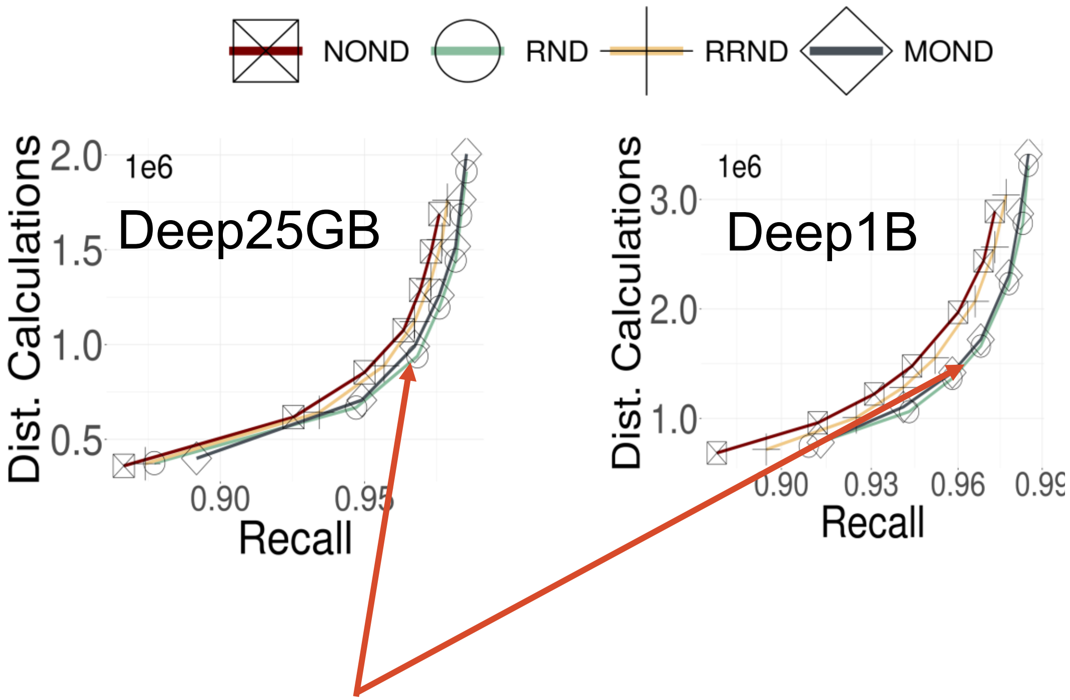


(b) Local Relative Contrast (LRC): high values indicate easy search

Fig. 4. Dataset Complexity

Experimental Evaluation - ND

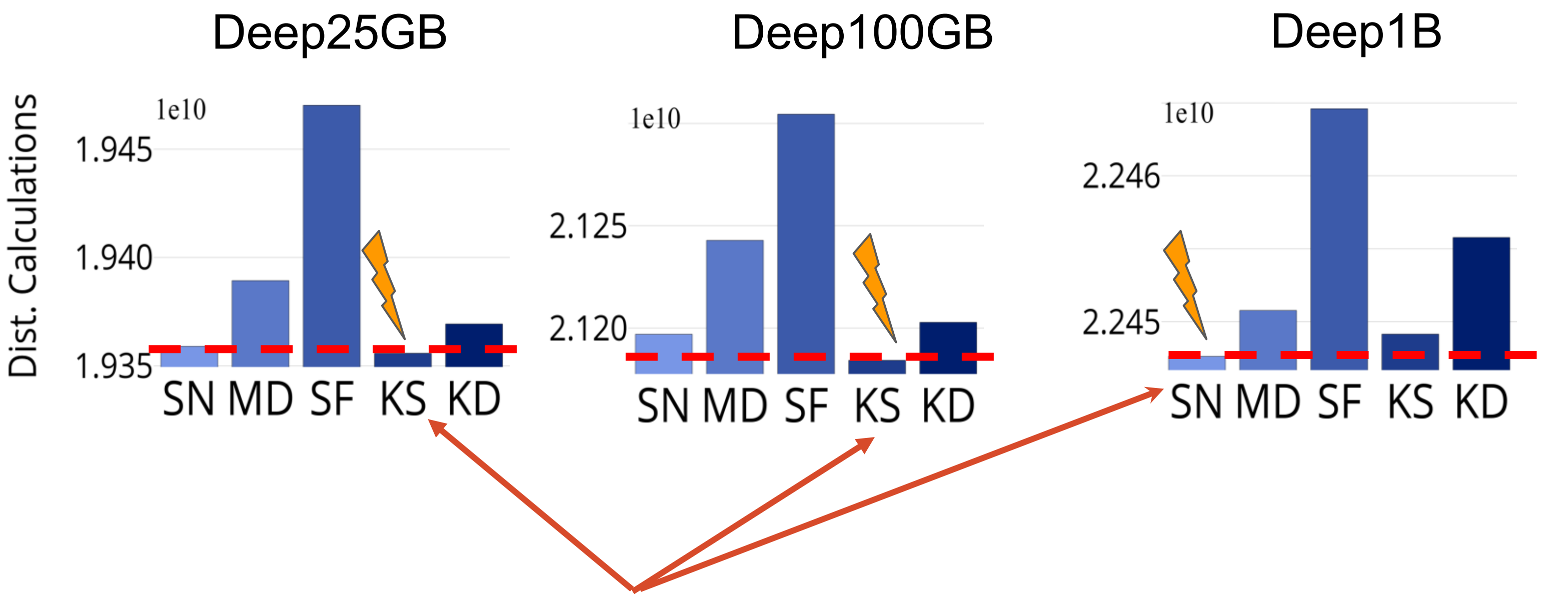


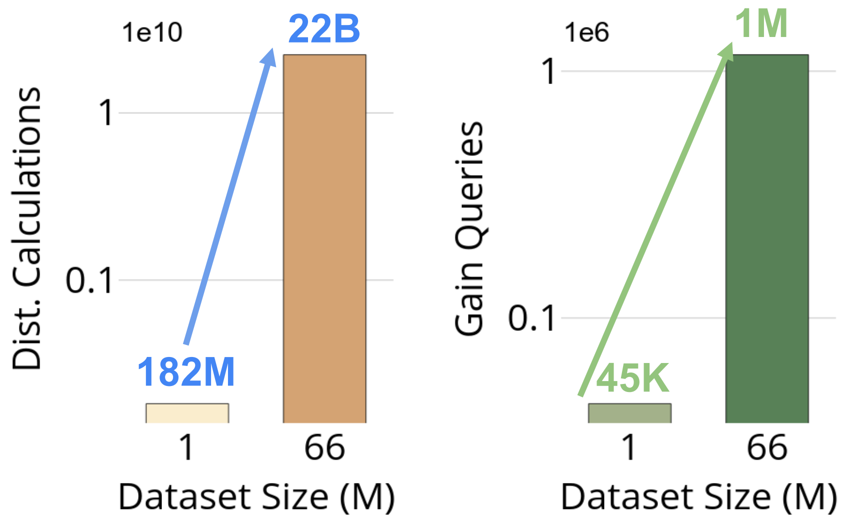


RND prunes at least **5x and 20x** more than **MOND** and **RRND, leading to a smaller index size** and **search memory footprint**.

RND leads to the best **search efficiency** across datasets and dataset sizes.

Experimental Evaluation - SS





Deep1M and Deep25GB
(SN-KS)

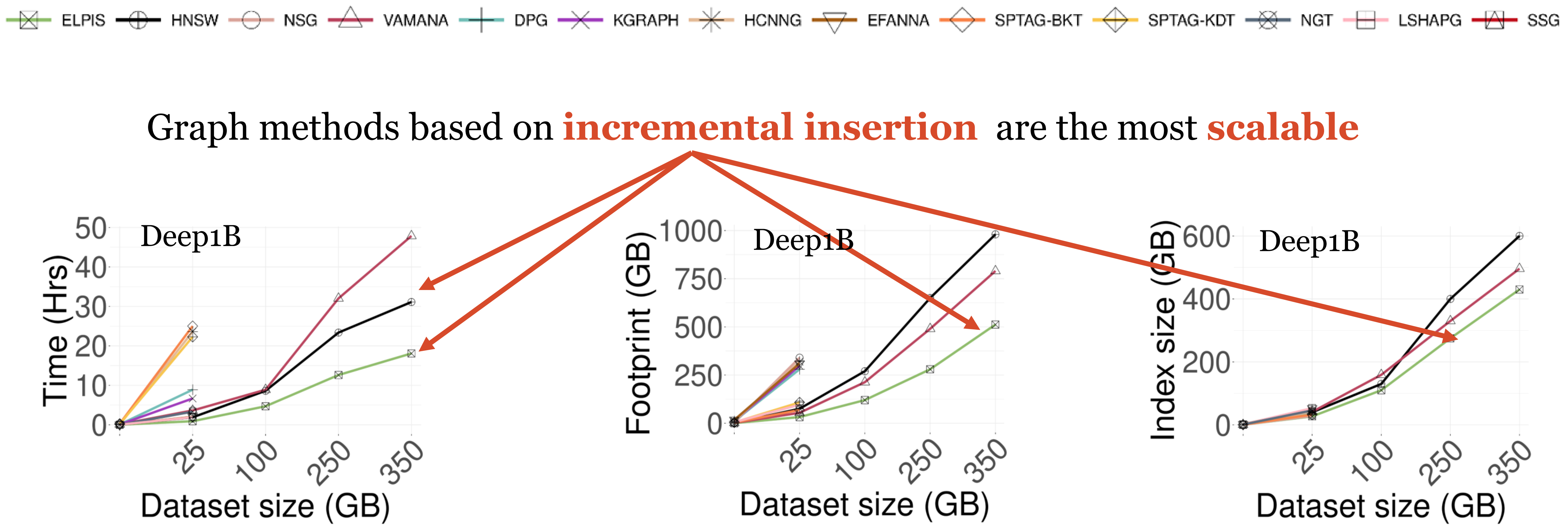
Optimizing **data structures** for **seed selection** enhances **search efficiency** on **large datasets**

The choice of seed selection impacts **index efficiency** as well

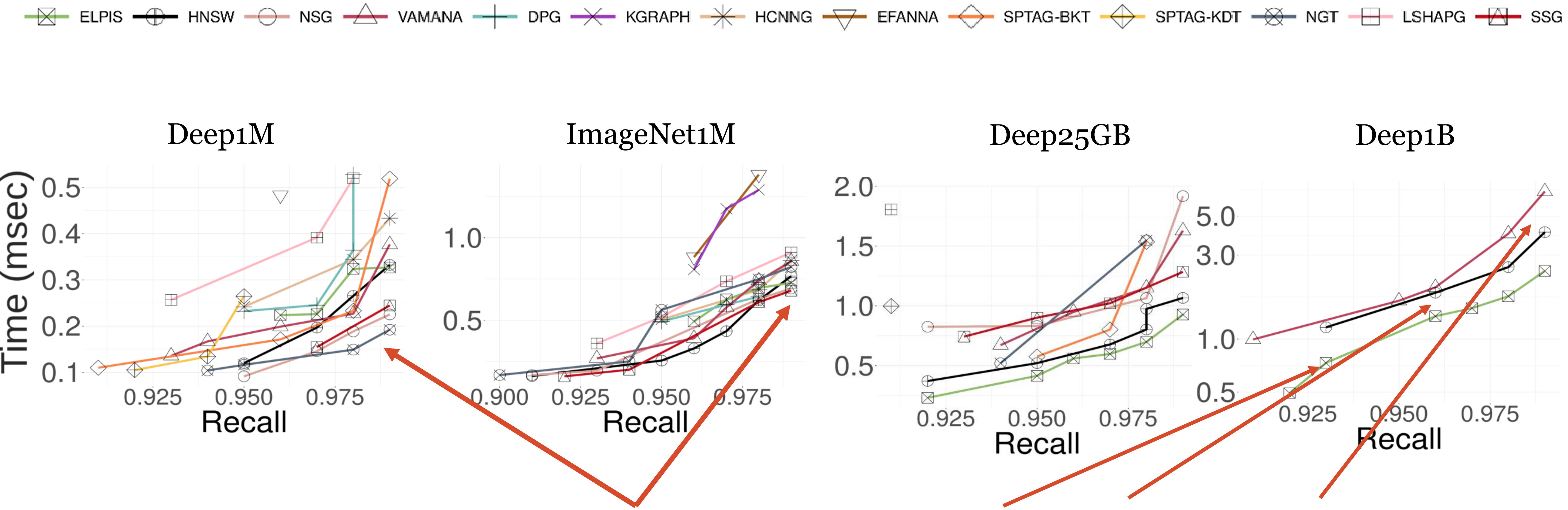
Experimental Evaluation - Baselines

- HNSW [12]
- NSG [10]
- VAMANA [13]
- SPTAG [14]
- NGT [15]
- SSG [9]
- LSH-APG [17]
- HCNNG [18]
- DPG [11]
- EFANNA [8]
- KGRAPH [7]
- ELPIS [X]

Experimental Evaluation – Indexing Performance



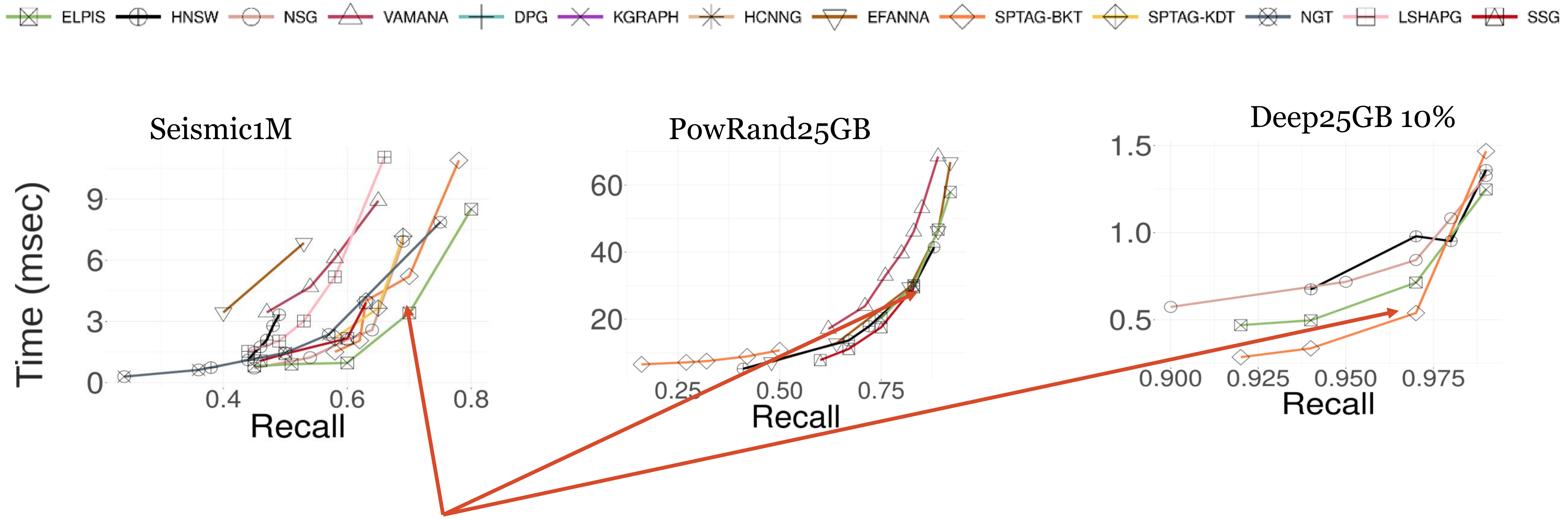
Experimental Evaluation – Search Performance



graph-based methods based on **ND** lead to the best **search performance**

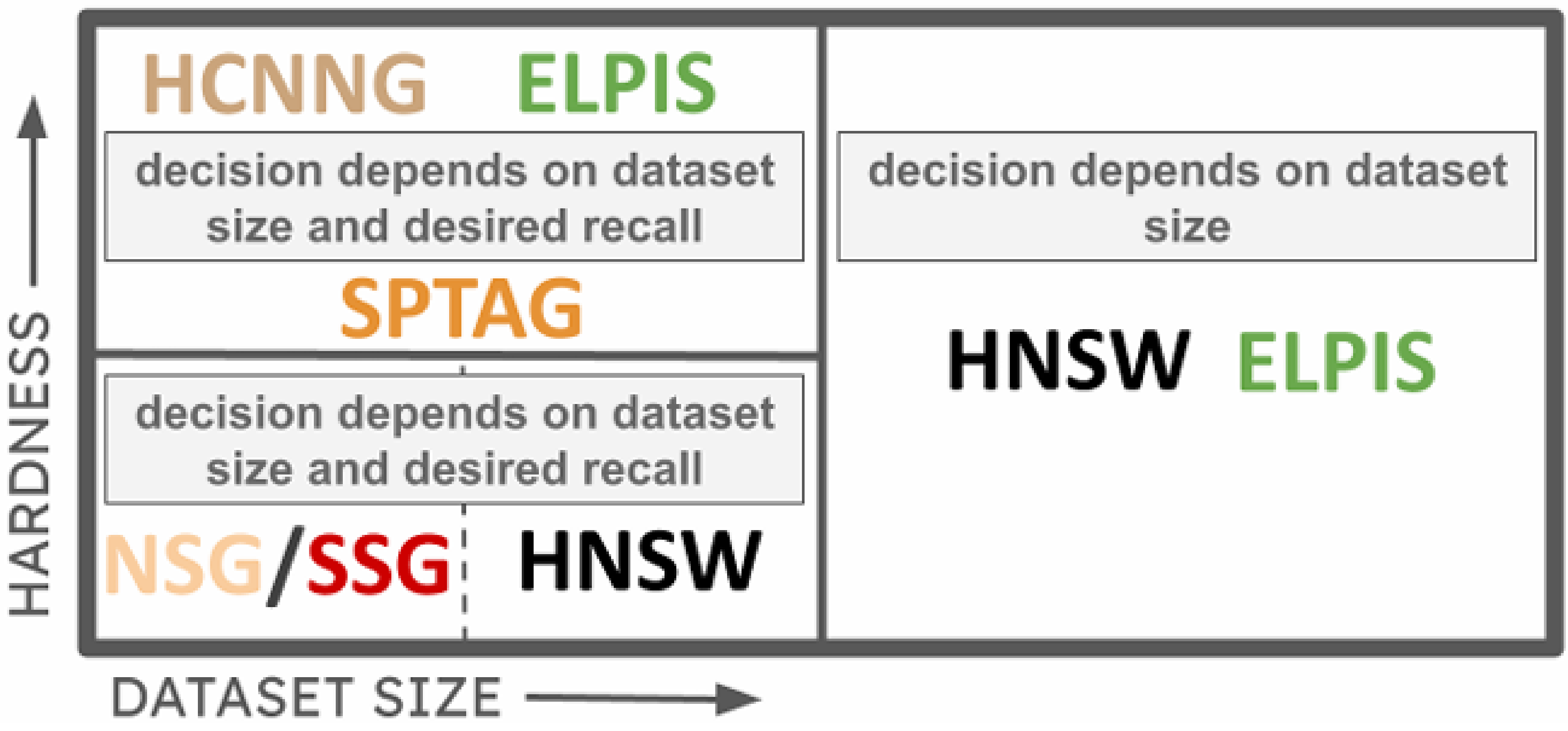
Elpis, HNSW scale in **billion-scale datasets**

Experimental Evaluation – Search Performance



Divide and Conquer graph-based methods lead the best **search performance on hard datasets and workloads**

Recommendations



<u>Indexing + 10K queries (0.99 recall)</u>

Unexpected Results

- Simple SS approaches like K-random sampling outperform Stacked NSW on small and medium sized datasets.
- Some methods (SPTAG, NGT, NSG, and SSG) cannot build an index on large datasets (>=100GB) within hours, despite excellent performance on smaller datasets (1M and 25GB).
- DC-based approaches (Elpis, SPTAG) outperform the others on challenging datasets/workloads.

Key Takeaways

- No method wins across the board.
  - II-based methods scale best overall at indexing and query answering
  - ND-based methods have superior query answering
  - DC-based methods scale best at indexing and challenging datasets/query workloads

- Adopting ND to sparsify the graph ***always*** leads to better search performance, particularly on large datasets.

- Effective SS plays a crucial role in enhancing both search and indexing performance.

Promising Research Directions

- Graph-based search:
  - Theoretical studies to better understand the trade-offs between proximity and sparsity.
  - Lightweight SS strategies to further improve search and indexing performance, particularly for out-of-distribution queries.
  - Hybrid methods that combines the strengths of II, ND and DC.
  - Adaptive methods that cater to dataset characteristics such as dataset size, dimensionality, RC and LID.
  - Novel base graphs, clustering and summarization techniques tailored for DC-based methods can further improve their performance.
  - Filtered search:
    - Range search
    - Updates
    - Non-ED distances
    - Disk-based

Promising Research Directions

- Tree-based search:
  - Improve summarization techniques for exact search
    - Higher tightness of the lower-bound
    - Cheaper to compute
  - Support probabilistic/deterministic approximate search
    - More effective stopping criteria
    - Guarantees on recall not only distance error
  - Exploit modern hardware

Open for collaboration!

visit: https://**echihabi.com**

Presented tutorials on the topic of vector search:
IEEE Big Data 2020, EDBT 2021, ICDE 2021, VLDB 2021, MDM 2022

Open for collaboration!

visit: [https://**echihabi.com**](https://echihabi.com)

Presented tutorials on the topic of vector search:
IEEE Big Data 2020, EDBT 2021, ICDE 2021, VLDB 2021, MDM 2022

vector search, data mining, data valuation, applications to RAG/GenAI pipelines and health

References (chronological order)

- G. Salton. Some hierarchical models for automatic document retrieval. *American Documentation*, *14*(3), 213-222, 1963.
- Jon Louis Bentley. 1975. Multidimensional binary search trees used for associative searching. Commun. ACM 18, 9 (Sept. 1975), 509–517.
- Guttman 'R-trees a dynamic index structure for spatial searching', Proc ACM SIGMOD Int Conf on Management of Data, 47-57, 1984
- Kenneth L. Clarkson. A randomized algorithm for closest-point queries. SIAM Journal on Computing, 17(4):830–847, 1988.
- Chin, Andrew (1994). Locality-Preserving Hash Functions for General Purpose Parallel Computation. Algorithmica. 12 (2–3)
- Gionis, A.; Indyk, P.; Motwani, R. (1999). "Similarity Search in High Dimensions via Hashing". Proceedings of the 25th Very Large Database (VLDB) Conference.
- K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "'nearest neighbor'" meaningful? In ICDT, 1999.

References (chronological order)

- C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional spaces. In ICDT, 2001.
- Samet, H. (2006). Foundations of multidimensional and metric data structures. Morgan Kaufmann.
- Stephen Blott and Roger Weber. 2008. What's wrong with high-dimensional similarity search? Proc. VLDB Endow. 1, 1 (August 2008), 3.
- Hervé Jégou, Matthijs Douze, Cordelia Schmid: Product Quantization for Nearest Neighbor Search. IEEE TPAMI 33(1): 117-128 (2011)
- J. He, S. Kumar, and S.-F. Chang. On the difficulty of nearest neighbor search. In ICML, 2012.
- Y. Wang, P. Wang, J. Pei, W. Wang, and S. Huang. A data-adaptive and dynamic segmentation index for whole matching on time series. PVLDB, 6(10):793–804, 2013.
- Alessandro Camerra, Jin Shieh, Themis Palpanas, Thanawin Rakthanmanon, Eamonn J. Keogh: Beyond one billion time series: indexing and mining very large time series collections with iSAX2+. Knowl. Inf. Syst. 39(1): 123-151 (2014).

References (chronological order)

- Y. Malkov, A. Ponomarenko, A. Logvinov, and V. Krylov. Approximate nearest neighbor algorithm based on navigable small world graphs. Information Systems (IS), 45:61 – 68, 2014.
- **Karima Echihabi**, Kostas Zoumpatianos, Themis Palpanas, Houda Benbrahim: The Lernaean Hydra of Data Series Similarity Search: An Experimental Evaluation of the State of the Art. Proc. VLDB Endow. 12(2): 112-127 (2018)
- **Karima Echihabi**, Kostas Zoumpatianos, Themis Palpanas, Houda Benbrahim: Return of the Lernaean Hydra: Experimental Evaluation of Data Series Approximate Similarity Search. Proc. VLDB Endow. 13(3): 403-420 (2019)
- S. J. Subramanya, R. Kadekodi, R. Krishaswamy, and H. V. Simhadri. Diskann: Fast accurate billion-point nearest neighbor search on a single node. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, pages 13766–13776, 2019.
- Anna Gogolou, Theophanis Tsandilas, **Karima Echihabi**, Anastasia Bezerianos, Themis Palpanas: Data Series Progressive Similarity Search with Probabilistic Quality Guarantees. SIGMOD Conference 2020: 1857-1873
- Jafari, O., Maurya, P., Nagarkar, P., Islam, K. M., & Crushev, C. (2021). A survey on locality sensitive hashing algorithms and their applications. arXiv preprint arXiv:2102.08942.

References (chronological order)

- **Karima Echihabi**, Panagiota Fatourou, Kostas Zoumpatianos, Themis Palpanas, Houda Benbrahim: Hercules Against Data Series Similarity Search. Proc. VLDB Endow. 15(10): 2005-2018 (2022).
- **Karima Echihabi**, Theophanis Tsandilas, Anna Gogolou, Anastasia Bezerianos, Themis Palpanas: ProS: data series progressive k-NN similarity search and classification with probabilistic quality guarantees. VLDB J. 32(4): 763-789 (2023).
- Ilias Azizi, **Karima Echihabi**, Themis Palpanas: Elpis: Graph-Based Similarity Search for Scalable Data Science. Proc. VLDB Endow. 16(6): 1548-1559 (2023).
- Liana Patel, Peter Kraft, Carlos Guestrin, and Matei Zaharia. 2024. ACORN: Performant and Predicate-Agnostic Search Over Vector Embeddings and Structured Data. Proc. ACM Manag. Data 2, 3, Article 120 (May 2024), 27 pages. doi:10.1145/3654923.
- Hnswlib- fast approximate nearest neighbor search. https://github.com/nmslib/hnswlib. Accessed 2025-06-26.
- FacebookAIResearch.[n.d.]. Searchinginasubsetofelements. https://github.com/facebookresearch/faiss/wiki/Setting search-parameters-for-one-query#searching-in-a-subset-of-elements. Accessed: 2024-06-26.
- Ilias Azizi, **Karima Echihabi**, Themis Palpanas: Graph-Based Vector Search: An Experimental Evaluation of the State-of-the-Art. Proc. ACM Manag. Data 3(1): 43:1-43:31 (2025).
- Anas Ait Aomar, **Karima Echihabi**, Marco Arnaboldi, Ioannis Alagiannis, Damien Hilloulin, Manal Cherkaoui: RWalks: Random Walks as Attribute Diffusers for Filtered Vector Search. Proc. ACM Manag. Data 3(3): 212:1-212:26 (2025).
- Khaoula Abdenouri, **Karima Echihabi**. A Scalable Tree-Based Index for Exact Vector Search: Submitted. (2025).