

High-Dimensional Vector Similarity Search

From Time Series to Deep Network Embeddings

Karima Echiabi¹ (3rd year Ph.D Student, Supervised by Themis Palpanas² and Houda Benbrahim¹)

¹ Mohammed V University, Morocco

² Université de Paris & French University Institute (IUF), France

1. Problem and Motivation

- **Similarity search** is core operation for many critical data processing tasks
 - data cleaning, data integration and data analytics
- **Similarity search finds objects in a collection close to a given query according to some definition of sameness**
 - Objects can be **data series**, **images**, **graphs**, **deep network embeddings**, etc.
- **Similarity search algorithms can either return **exact** or **approximate** answers**
 - Exact methods return correct results but are expensive
 - Approximate methods sacrifice accuracy to achieve better efficiency
- 1) **no exact method is overall winner**
- 2) **no approximate method offers query accuracy guarantees, or scales to TBs of disk-based data** (with low footprint)

HERCULES: versatile similarity search

- **supports exact / approximate search**
- **best in accuracy / time / footprint**
- outperforms data series indexes, LSH, inverted indexes, kNNs on most scenarios

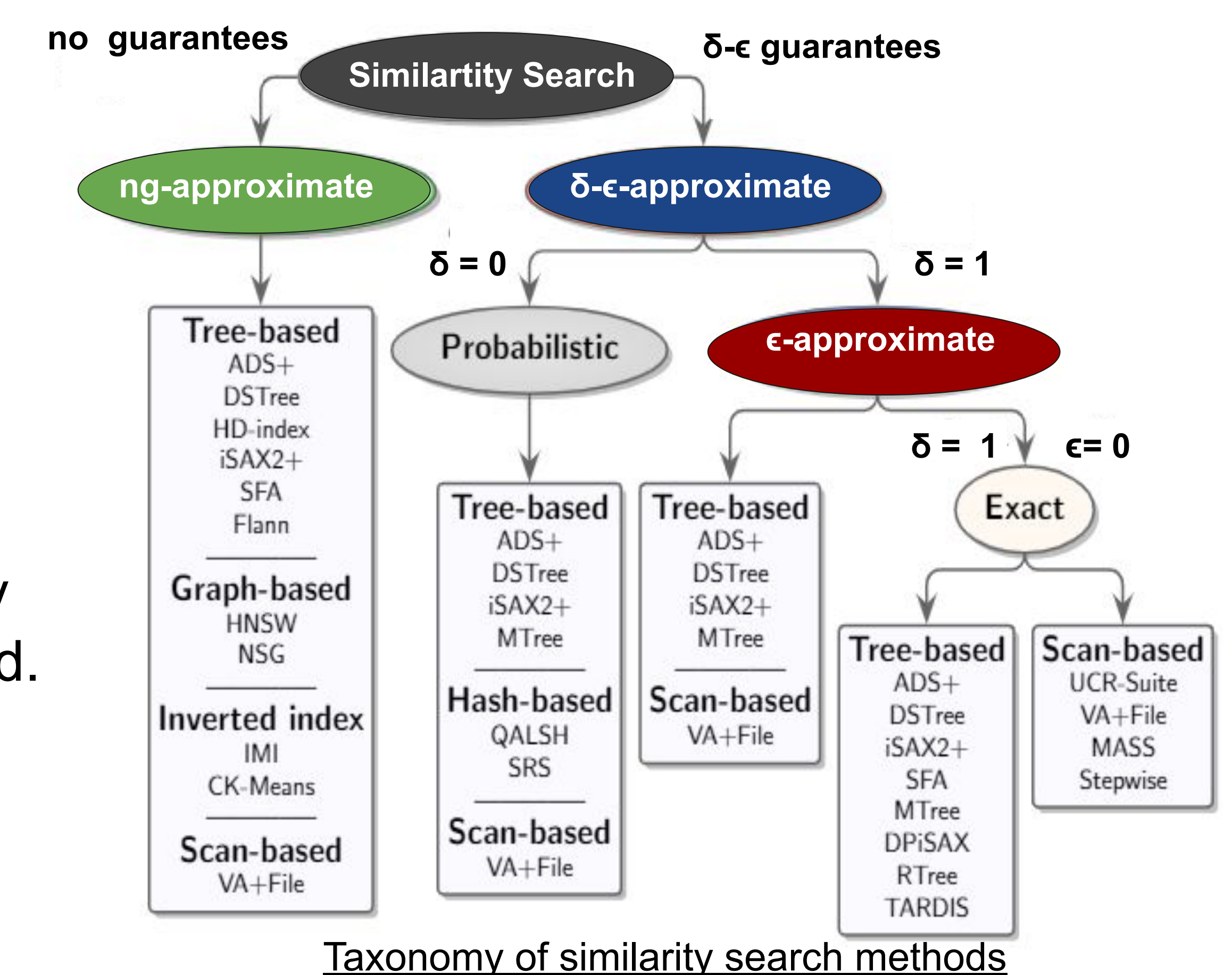
- **We conducted two most extensive **experimental evaluations** in this area**
 - Lernaean Hydra papers: exact (PVLDB'19) and approximate techniques (PVLDB'20)
 - considered **techniques from both data series and high-d vector communities**
- **Proposed extensions to existing indexes to **support δ - ϵ -approximate search****
 - **Outperform** LSH, kNNs and inverted indexes under many scenarios
- **Developed **Hercules**: novel versatile similarity search method**
 - Builds upon insights gained from Lernaean Hydra papers about intricate designs of existing approaches, their strengths and weaknesses
 - Supports **all flavors of exact and approximate search**
 - Supports **progressive query answering**

3. Approach and Uniqueness

2. Background and Related Work

- **Similarity search has been extensively studied in past 25 years**

- **Different communities** (data series and high-d vectors) as kNN problem)
- **Data series are special type of high-d vector data with ordered/correlated dimensions**
- kNN is still meaningful for many real data sets that are rarely i.i.d.



Taxonomy of similarity search methods

- **A deeper understanding of high-d vector exact and approximate similarity search**

- **A research direction that is more promising than the two current trends**
 - LSH-based algorithms, which **provide guarantees, but are slow**
 - kNNs and inverted indexes, which **are fast but provide no guarantees**
- **Hercules: novel similarity search method for exact/approximate search**
 - Current Prototype**
 - **Exact search: **Best overall****. At least 2x faster than best index, 20x-100x faster than Optimized Sequential Scan (OSS) on balanced workloads, only index that outperforms OSS on hard workloads
 - **Approximate search: **Best overall on disk**** (time/accuracy) with low footprint
- **A public archive of code/datasets/results**, a stepping stone for a benchmark

4. Results and Contributions