Scalable Analytics on Large Sequence Collections

Karima Echihabi
Mohammed VI Polytechnic University
karima.echihabi@um6p.ma

Themis Palpanas
LIPADE, Université Paris Cité & French University Institute (IUF)
themis@mi.parisdescartes.fr

***Abstract*—Data series are a prevalent data type that has attracted lots of interest in recent years. Specifically, there has been an explosive interest towards the analysis of large volumes of data series in many different domains, and in particular, in the Internet of Things (IoT). In this tutorial, we focus on applications that produce massive collections of data series, and we provide the necessary background on data series management and analytics. Moreover, we discuss the need for fast similarity search for supporting machine learning applications, and describe efficient similarity search techniques, indexes and query processing algorithms. Finally, we discuss the role that deep learning techniques can play in this context. We conclude with the challenges and open research problems in this domain.**

I. Introduction

In various scientific and industrial domains analysts are required to measure quantities as they fluctuate over a dimension; these values are commonly called *data series* or *sequences*. The dimension over which data series are ordered depends on the application domain and can have various diverse physical meanings. By far, the most common dimension over which data are ordered is time. In this case, we specifically talk about *time series*. Other applications though, produce series ordered over position (DNA sequences), mass (mass spectrometry) or angle (shapes). In all cases, data have to be processed as series rather than individual values.

Applications range from forecasting methods to correlation analysis, summarization, representation methods, outlier detection and more [5]–[10], [41], [50], [51], [53]. Recent advances in domains such as Internet of Things (IoT) and smart cities, self-driving cars and communications, generate tremendous amounts of data series, and drive the need for novel data series management solutions (Figure 1). Moreover, it is not unusual for applications to involve numbers of sequences in the order of hundreds of millions to billions [1], [3]. These data have to be analyzed, in order to identify patterns, gain insights, and detect abnormalities. As a result, analysts are more frequently than ever deluged by the vast amounts of data series that they have to filter, process and understand.

The goal of this tutorial is to describe the current state in data series management, including applications, data types, query types and complex analytic algorithms. Further on, we will explore how modern techniques can be leveraged to speed up complex analytical pipelines, and take a glimpse on how these techniques can be improved by applying machine learning.

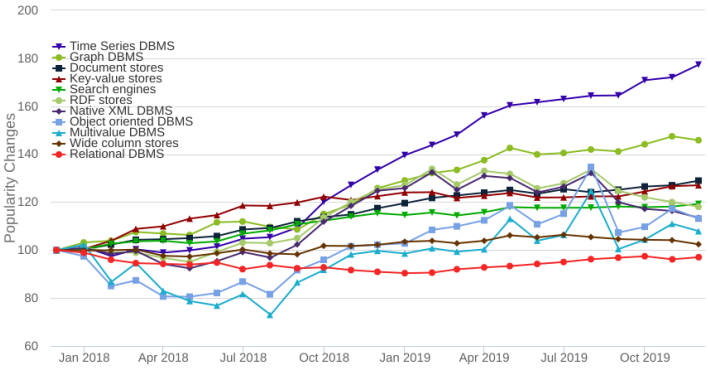

Fig. 1. Data management system category popularity change [2]

II. Projected Audience and Expected Background

Modern day research has entered an age where the amount of recorded data is increasing exponentially. The analysis of the time-series data associated with multiple fields is now beginning to push both computational power and resources to their limit. This tutorial is for both data analysts and researchers, and will focus on recent advances in academia and industry in the area of data series analysis. The tutorial aims at fostering collaborations between the data management community and data science practitioners in various domains, as well as on gathering awareness and interest on the domain of data series analytics.

In the material that we will present, we will include the background necessary in order to follow the entire presentation, as well as technical details of existing solutions, and discussions of drawbacks, open problems and challenges. Therefore, both experts and newcomers in the area will be able to follow the material and benefit from the tutorial.

III. Tutorial Scope

In this **1.5 hours** tutorial, we take a holistic look at the problem of managing and analyzing very large collections of data series, discuss the state-of-the-art and pinpoint the opportunities for optimizing complex query execution.

[Introduction and Foundations] We will start by looking at some foundational aspects of data series management. Those include the *data characteristics*, the query workloads, and the *specialized data structures* used to index sequential data. Data series can be categorized under many dimensions: the way that data arrive (streaming vs static), the lengths of data

series (fixed vs variable length per series), the way that points are sampled (fixed intervals vs variable sampling intervals), and the presence of uncertainty in their values. In terms of workloads, we will then look at various applications and query patterns that recur in each one of those. Specifically, we will discuss both simple Selection-Projection-Transformation (SPT) queries, where analysts filter based on data properties (e.g., thresholds) or meta-data values, as well as complex data mining (DM) analytics, like clustering, outlier detection and more [48].

We will look at the core component of advanced analytics, which is similarity search, and look at the different flavors of this problem. Those include whole matching vs sub-sequence matching, exact vs approximate similarity search, as well as various distance measures that are commonly used in practice. Finally, we will briefly talk about the different data structure categories that exist, and how they are used to organize and retrieve data in each one of the aforementioned query patterns.
[Complex Analytics] We will dive in analytics like outlier detection [12], [17], frequent pattern mining [61], clustering [32], [62], [63], [72], and classification [16]. Such analytics involve a series of operations that are performed in a pre-processing step as well as operations that are repeated in the context of an iterative algorithm. Pre-processing operations include sliding windows, normalization, interpolation, and various transformations such as DFT that are specific to each algorithm.

During the iterative part of these analytics, multiple similarity search operations need to be performed. This is useful for finding series within a given radius from a centroid in clustering, or for identifying distances from a given model in anomaly detection and classification, but also for retrieving patterns in frequent pattern mining. All of these operations can be implemented externally, in the application side. However, since some of them are data-intensive, pruning or incremental computation can significantly improve their performance. For this reason, performing them at the database level can provide large improvements in terms of execution time. We will focus on similarity search as such an example, being a crucial and expensive component of most mining algorithms, and motivate a deep-dive at its characteristics and scalable implementations.
[Advanced Techniques for Optimizing Analytics] We will present techniques for speeding up similarity search, which plays a central role in several algorithms related to complex data series analytics. Previous work on similarity search has proposed the use of spatial indexes such as R-Trees with DFT [4], [60] and DHWT [11]. Specialized indexes are based on domain specific summarizations. Examples include DS-Tree [70], iSAX [49], [67], ADS [78], SFA [66], Coconut [33], [34], KV-Match [73], L-Match [29], TS-Index [15], ULISSE [39], [40], DPiSAX [74], [75], TARDIS [76], ParIS+ [55], [57], MESSI [56], [58], SING [59], and Hercules [21]. Moreover, specialized techniques have been developed for geolocated data series [13], [14]. Recent studies [19], [20], [27], [28], [38] have compared several data series and high-dimensional similarity search methods under a common framework, revealing multiple promising future research directions, which we will analyze.

We will also discuss how **deep learning advances** can be leveraged to push the frontiers of big sequence management We will review techniques used for data series and cover recent successes in related areas. We identify three main directions: (1) learning accurate and concise summarizations; (2) designing efficient data structures; and (3) performing query optimization. Specific methods for data series have also been proposed for learning features [77] and embeddings [37], [52], [54], [69], or encoding data series as images to leverage computer vision techniques [71]. A recent paper [36] introduced the idea of using machine learning techniques to build a new class of indexes for one-dimensional data, and other works extended this notion to multidimensional data [18], [42], [46], [64]. The development of learned data structures for data series is still an open question. Access path selection for data series is another promising research direction [27]. Since this field is still at its infancy, ideas can be borrowed from techniques and approaches based on deep learning [35], [43], [47].
[Challenges and Conclusions] Massive data series collections are becoming a reality for virtually every scientific and social domain. This leads to the need of designing and developing general-purpose Data Series Management Systems, able to cope with big data series, that is, very large and fast-changing collections of data series, which can be heterogeneous (i.e., originate from disparate domains and thus exhibit very different characteristics), and which can have uncertainty in their values (e.g., due to inherent errors in the measurements). These systems should have data series indexes and summarizations integrated into their engines, so as to speedup the time-intensive operations of complex analytics pipelines, and support interactive exploration of big data series. To this end, progressive analytics operators would also be very useful [30], [31], [68]. At the same time, the role that deep learning techniques can play should be studied in more detail, especially with regards to similarity search [24] and query optimization. Finally, there is a pressing need for developing data series specific benchmarks able to stress test index structures [79], [80] and other analysis tasks [53] in a principled way.

A. Outline

Next, we report the outline of the tutorial (duration: 1.5h).

1) **Introduction, Motivation, and Foundations (15 min)**
   - Data series domains and application examples
   - Data series and their properties (streaming/static, fixed/variable length, fixed/arbitrary sampling rate, univariate/multivariate, etc.)
   - Different types of data series queries (simple queries/complex analytics, subsequence-/whole-matching, approximate/progressive/exact)
2) **Advanced Techniques for Analytics (60 min)**
   - Summarizing and compressing data series
   - Similarity search: Exact and approximate methods (iSAX2+, ADS, DS-Tree, SFA, Coconut, KV-Match, L-Match, TS-Index, ULISSE, DPiSAX,

TARDIS, ParIS+, MESSI, SING, Hercules, etc.) and alternatives (scan acceleration/progressive search)
- Deep learning for data series summarization and embeddings, indexing and similarity search
- Results and lessons from extensive experimental comparisons of similarity search methods

3) **Challenges and Conclusions (15 min)**
- Open problems in data series management systems
- Open problems in complex data series analytics
- Opportunities and challenges for deep learning techniques

Relation to Previous Tutorials. We note that previous tutorials in the domain of data series have either concentrated on time series mining algorithms [45], [65], or specifically on the characteristics of different time series similarity measures [44], with no reference to index data structures, which are now becoming popular and necessary for handling the increasingly larger data series collections in different applications across many domains.

Compared to previous versions [22], [23], [25], [26] of this tutorial, we now include several new techniques that we will present in more detail, we will discuss approaches that handle geolocated data series, and we will explore in more depth the opportunities that deep learning has to offer.

IV. Presenters

Karima Echihabi is an assistant professor of computer science at the Mohammed VI Polytechnic University (Morocco). Her research interests lie in data analytics and data series. She has performed extensive analysis of data series indexes and has delivered 5 tutorials in top international conferences.

Themis Palpanas is a Senior Member of the French University Insitute (IUF), and a professor of computer science at the Université Paris Cité (France). He has been working in the fields of Data Series Management and Analytics for more than 15 years, and has developed several of the state of the art techniques in major journals and conferences. He has delivered 15 tutorials in top international conferences, including data management (VLDB/SIGMOD), information retrieval (SIGIR), and the Web (WWW).

Acknowledgments. Work supported by program Investir l'Avenir and University of Paris IDEX Emergence en Recherche ANR-18-IDEX-0001, EU project NESTOR (MSCA #748945), FMJH Program PGMO in conjunction with EDF and THALES, and Institut Universitaire de France (IUT).

References

[1] Adhd-200. http://fcon_1000.projects.nitrc.org/indi/adhd200/.
[2] Db-engines. https://db-engines.com/en/ranking_categories.
[3] Sloan digital sky survey. https://www.sdss3.org/dr10/data_access/rvolume.php.
[4] R. Agrawal, C. Faloutsos, and A. N. Swami. Efficient similarity search in sequence databases. In *FODO*, 1993.
[5] A. J. Bagnall, R. L. Cole, T. Palpanas, and K. Zoumpatianos. Data series management (dagstuhl seminar 19282). *Dagstuhl Reports*, 9(7), 2019.
[6] P. Boniol, M. Linardi, F. Roncallo, and T. Palpanas. Automated anomaly detection in large sequences. In *ICDE*, 2020.
[7] P. Boniol, M. Linardi, F. Roncallo, T. Palpanas, M. Meftah, and E. Remy. Unsupervised and scalable subsequence anomaly detection in large data series. *VLDBJ*, 2021.
[8] P. Boniol, M. Meftah, E. Remy, and T. Palpanas. dCAM: Dimension-wise Activation Map for Explaining Multivariate Data Series Classification. In *SIGMOD*, 2022.
[9] P. Boniol and T. Palpanas. Series2Graph: Graph-based Subsequence Anomaly Detection for Time Series. *PVLDB*, 2020.
[10] P. Boniol, J. Paparrizos, T. Palpanas, and M. J. Franklin. SAND: streaming subsequence anomaly detection. *PVLDB*, 2021.
[11] K. Chan and A. W. Fu. Efficient time series matching by wavelets. In *ICDE*, 1999.
[12] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.
[13] G. Chatzigeorgakidis, D. Skoutas, K. Patroumpas, T. Palpanas, S. Athanasiou, and S. Skiadopoulos. Local Pair and Bundle Discovery over Co-Evolving Time Series. In *SSTD*, 2019.
[14] G. Chatzigeorgakidis, D. Skoutas, K. Patroumpas, T. Palpanas, S. Athanasiou, and S. Skiadopoulos. Local Similarity Search on Geolocated Time Series Using Hybrid Indexing. In *SIGSPATIAL*, 2019.
[15] G. Chatzigeorgakidis, D. Skoutas, K. Patroumpas, T. Palpanas, S. Athanasiou, and S. Skiadopoulos. Efficient Range and kNN Twin Subsequence Search in Time Series. *TKDE*, 2022.
[16] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *JMLR*, 2009.
[17] M. Dallachiesa, T. Palpanas, and I. F. Ilyas. Top-k nearest neighbor search in uncertain data series. *PVLDB*, 8(1), 2014.
[18] Y. Dong, P. Indyk, I. P. Razenshteyn, and T. Wagner. Learning sublinear-time indexing for nearest neighbor search. *CoRR*, abs/1901.08544, 2019.
[19] K. Echihabi. Truly Scalable Data Series Similarity Search. In *Proceedings of the VLDB 2019 PhD Workshop*, 2019.
[20] K. Echihabi. High-dimensional vector similarity search: From time series to deep network embeddings. In *SIGMOD*, 2020.
[21] K. Echihabi, P. Fatourou, K. Zoumpatianos, T. Palpanas, and H. Benbrahim. Hercules Against Data Series Similarity Search. *PVLDB*, 2022.
[22] K. Echihabi, T. Palpanas, and K. Zoumpatianos. New trends in high-d vector similarity search: Ai-driven, progressive, and distributed. *PVLDB*, 2021.
[23] K. Echihabi, K. Zoumpatianos, and T. Palpanas. Big sequence management: on scalability. In *IEEE BigData*, 2020.
[24] K. Echihabi, K. Zoumpatianos, and T. Palpanas. Scalable Machine Learning on High-Dimensional Vectors: From Data Series to Deep Network Embeddings. In *WIMS*, 2020.
[25] K. Echihabi, K. Zoumpatianos, and T. Palpanas. Big sequence management: Scaling up and out. In *EDBT*, 2021.
[26] K. Echihabi, K. Zoumpatianos, and T. Palpanas. High-dimensional similarity search for scalable data science. ICDE, 2021.
[27] K. Echihabi, K. Zoumpatianos, T. Palpanas, and H. Benbrahim. The lernaean hydra of data series similarity search: An experimental evaluation of the state of the art. *PVLDB*, 12(2), 2018.
[28] K. Echihabi, K. Zoumpatianos, T. Palpanas, and H. Benbrahim. Return of the Lernaean Hydra: Experimental Evaluation of Data Series Approximate Similarity Search. *PVLDB*, 2019.
[29] K. Feng, P. Wang, J. Wu, and W. Wang. L-match: A lightweight and effective subsequence matching approach. *IEEE Access*, 2020.
[30] A. Gogolou, T. Tsandilas, K. Echihabi, T. Palpanas, and A. Bezerianos. Data Series Progressive Similarity Search with Probabilistic Quality Guarantees. In *SIGMOD*, 2020.
[31] A. Gogolou, T. Tsandilas, T. Palpanas, and A. Bezerianos. Progressive similarity search on time series data. In *EDBT*, 2019.

[32] E. Keogh and M. Pazzani. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In *KDD*, 1998.

[33] H. Kondylakis, N. Dayan, K. Zoumpatianos, and T. Palpanas. Coconut: A scalable bottom-up approach for building data series indexes. *PVLDB*, 2018.

[34] H. Kondylakis, N. Dayan, K. Zoumpatianos, and T. Palpanas. Coconut: sortable summarizations for scalable indexes over static and streaming data series. *VLDBJ*, 2019.

[35] T. Kraska, M. Alizadeh, A. Beutel, E. H. Chi, J. Ding, A. Kristo, G. Leclerc, S. Madden, H. Mao, and V. Nathan. Sagedb: A learned database system. In *CIDR*, 2019.

[36] T. Kraska, A. Beutel, E. H. Chi, J. Dean, and N. Polyzotis. The case for learned index structures. In *SIGMOD*, 2018.

[37] Q. Lei, J. Yi, R. Vaculín, L. Wu, and I. S. Dhillon. Similarity preserving representation learning for time series analysis. *CoRR*, abs/1702.03584, 2017.

[38] O. Levchenko, B. Kolev, D. E. Yagoubi, R. Akbarinia, F. Masseglia, T. Palpanas, D. E. Shasha, and P. Valduriez. BestNeighbor: Efficient Evaluation of kNN Queries on Large Time Series Databases. *Knowl. Inf. Syst.*, 63(2):349–378, 2021.

[39] M. Linardi and T. Palpanas. Scalable, variable-length similarity search in data series: The ULISSE approach. *PVLDB*, 11(13):2236–2248, 2018.

[40] M. Linardi and T. Palpanas. Scalable data series subsequence matching with ULISSE. *VLDB J.*, 29(6):1449–1474, 2020.

[41] M. Linardi, Y. Zhu, T. Palpanas, and E. J. Keogh. Matrix Profile Goes MAD: Variable-Length Motif And Discord Discovery in Data Series. 2020.

[42] S. Macke, A. Beutel, T. Kraska, M. Sathiamoorthy, D. Zhiyuan Cheng, and C. E. H. Lifting the curse of multidimensional data with learned existence indexes. In *ML for Systems Workshop at NIPS*, 2018.

[43] R. Marcus and O. Papaemmanouil. Towards a hands-free query optimizer through deep learning. In *CIDR*, 2019.

[44] A. Mueen. Similarity search on time series data: Past, present, and future. In *CIKM*, 2016.

[45] A. Mueen and E. Keogh. Finding repeated structure in time series: Algorithms and applications. In *ICDM*, 2014.

[46] V. Nathan, J. Ding, M. Alizadeh, and T. Kraska. Learning multi-dimensional indexes. In *SIGMOD*, 2019.

[47] J. Ortiz, M. Balazinska, J. Gehrke, and S. S. Keerthi. Learning state representations for query optimization with deep reinforcement learning. In *DEEM@SIGMOD*, pages 4:1–4:4, 2018.

[48] T. Palpanas. Data series management: The road to big sequence analytics. *SIGMOD Rec.*, 44(2):47–52, 2015.

[49] T. Palpanas. Evolution of a Data Series Index. *CCIS*, 1197, 2020.

[50] T. Palpanas and V. Beckmann. Report on the first and second interdisciplinary time series analysis workshop (itisa). *SIGREC*, 48(3), 2019.

[51] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. Volume Under the Surface: A New Accuracy Evaluation Measure for Time-Series Anomaly Detection. *Technical Report LIPADE-TR-N7, Université Paris Cité*, 2022.

[52] J. Paparrizos and M. J. Franklin. GRAIL: efficient time-series representation learning. *PVLDB*, 12(11):1762–1777, 2019.

[53] J. Paparrizos, Y. Kang, P. Boniol, R. S. Tsay, T. Palpanas, and M. J. Franklin. TSB-UAD: An End-to-End Benchmark Suite for Univariate Time-Series Anomaly Detection. *PVLDB*, 2022.

[54] W. Pei, D. M. J. Tax, and L. van der Maaten. Modeling time series similarity with siamese recurrent networks. *CoRR*, abs/1603.04713, 2016.

[55] B. Peng, P. Fatourou, and T. Palpanas. ParIS: The Next Destination for Fast Data Series Indexing and Query Answering. In *IEEE BigData*, 2018.

[56] B. Peng, P. Fatourou, and T. Palpanas. Messi: In-memory data series indexing. In *ICDE*, 2020.

[57] B. Peng, P. Fatourou, and T. Palpanas. Paris+: Data series indexing on multi-core architectures. *TKDE*, 2020.

[58] B. Peng, P. Fatourou, and T. Palpanas. Fast data series indexing for in-memory data. *VLDB J.*, 30(6):1041–1067, 2021.

[59] B. Peng, P. Fatourou, and T. Palpanas. SING: Sequence Indexing Using GPUs. In *ICDE*, 2021.

[60] D. Rafiei and A. O. Mendelzon. Similarity-based queries for time series data. In *SIGMOD*, 1997.

[61] T. Rakthanmanon, B. J. L. Campana, A. Mueen, G. Batista, M. B. Westover, Q. Zhu, J. Zakaria, and E. J. Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In *KDD*, 2012.

[62] T. Rakthanmanon, E. J. Keogh, S. Lonardi, and S. Evans. Time series epenthesis: Clustering time series streams requires ignoring some data. In *ICDM*, 2011.

[63] P. P. Rodrigues, J. Gama, and J. P. Pedroso. Odac: Hierarchical clustering of time series data streams. In *SDM*, 2006.

[64] A. Sablayrolles, M. Douze, C. Schmid, and H. Jégou. Spreading vectors for similarity search. *ICLR*, 2019.

[65] Y. Sakurai, Y. Matsubara, and C. Faloutsos. Mining big time-series data on the web. In *WWW*, 2016.

[66] P. Schäfer and M. Högqvist. SFA: a symbolic fourier approximation and index for similarity search in high dimensional datasets. In *EDBT*, 2012.

[67] J. Shieh and E. J. Keogh. *i*sax: indexing and mining terabyte sized time series. In *KDD*, 2008.

[68] C. Turkay, N. Pezzotti, C. Binnig, H. Strobelt, B. Hammer, D. A. Keim, J. Fekete, T. Palpanas, Y. Wang, and F. Rusu. Progressive data science: Potential and challenges. *CoRR*, abs/1812.08032, 2018.

[69] Q. Wang and T. Palpanas. Deep learning embeddings for data series similarity search. In *SIGKDD*, 2021.

[70] 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.

[71] Z. Wang and T. Oates. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In *Workshops at AAAI*, 2015.

[72] T. Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[73] J. Wu, P. Wang, N. Pan, C. Wang, W. Wang, and J. Wang. Kv-match: A subsequence matching approach supporting normalization and time warping. In *ICDE*, 2019.

[74] D. E. Yagoubi, R. Akbarinia, F. Masseglia, and T. Palpanas. DPiSAX: Massively Distributed Partitioned iSAX. In *ICDM*, 2017.

[75] D.-E. Yagoubi, R. Akbarinia, F. Masseglia, and T. Palpanas. Massively distributed time series indexing and querying. *TKDE*, 32(1), 2020.

[76] L. Zhang, N. Alghamdi, M. Y. Eltabakh, and E. A. Rundensteiner. TARDIS: distributed indexing framework for big time series data. In *ICDE*, 2019.

[77] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao. Time series classification using multi-channels deep convolutional neural networks. In *WAIM*, pages 298–310. Springer, 2014.

[78] K. Zoumpatianos, S. Idreos, and T. Palpanas. ADS: the adaptive data series index. *VLDB J.*, 25(6):843–866, 2016.

[79] K. Zoumpatianos, Y. Lou, I. Ileana, T. Palpanas, and J. Gehrke. Generating data series query workloads. *VLDB J.*, 27(6), 2018.

[80] K. Zoumpatianos, Y. Lou, T. Palpanas, and J. Gehrke. Query workloads for data series indexes. In *KDD*, 2015.