Report on the First and Second Interdisciplinary Time Series Analysis Workshop (ITISA)

Themis Palpanas
Paris Descartes University
themis@mi.parisdescartes.fr

Volker Beckmann
CNRS, Paris Diderot University
beckmann@in2p3.fr

ABSTRACT

The analysis of time-series data associated with modern-day industrial operations and scientific experiments is now pushing both computational power and resources to their limits. In order to analyze the existing and (more importantly) future very large time series collections, new technologies and the development of more efficient and smarter algorithms are required. The two editions of the Interdisciplinary Time Series Analysis Workshop brought together data analysts from the fields of computer science, astrophysics, neuroscience, engineering, electricity networks, and music. The focus of these workshops was on the requirements of different applications in the various domains, and also on the advances in both academia and industry, in the areas of time-series management and analysis. In this paper, we summarize the experiences presented in and the results obtained from the two workshops, highlighting the relevant state-of-the-art-techniques and open research problems.

1. INTRODUCTION

Time series[1] have gathered the attention of the data management community for more than two decades [1, 15, 30]. They are one of the most common data types, present in virtually every scientific and social domain: they appear as audio sequences [13], shape and image data [29], financial [26], environmental monitoring [25] and scientific data [11], and they have many diverse applications, such as in health care, astronomy, biology, economics, etc.

Recent advances in sensing, networking, data processing and storage technologies have significantly eased the process of generating and collecting data series. It is not unusual for applications to involve numbers of sequences in the order of hundreds of millions to billions [21]. These data have to be analyzed to identify patterns, gain insights, detect anomalies, and extract new knowledge.

A key observation is that analysts need to process a sequence (or subsequence) of values as a single object, rather than the individual points independently, which is what makes the management and analysis of data sequences a hard problem. Note that even though a sequence can be regarded as a point in n-dimensional space, traditional multi-dimensional approaches fail in this case, mainly due to the combination of the following two reasons: (a) the dimensionality is typically very high, i.e., in the order of several hundreds to several thousands, and (b) dimensions are strictly ordered (imposed by the sequence) and neighboring values are correlated.

Current time series analysis solutions require custom code, which implies huge investments in time and effort, and duplication of effort across different teams. Existing systems (e.g., based on DBMSs, Column Stores, or Array Databases) do not provide a viable solution, since they have not been designed for managing and processing sequence data [21]. Therefore, they do not offer a suitable declarative query language, storage model, auxiliary data structures, and optimization mechanism that can support a variety of sequence query workloads in an efficient manner [6,31]. (In Section 4, we discuss more reasons why existing solutions are inadequate.)

The Interdisciplinary Time Series Analysis Workshop provided a forum for researchers and practitioners that approach time series from different angles, ranging from data management and processing, to analysis, mining and machine learning. The core research issues considered in the workshop include: management and indexing, interactive visualization, machine learning, privacy preserving analytics, uncertainty and missing values, and applications of those in astrophysics, neuroscience, engineering, electricity networks, and music.

The program of the two editions of the workshop included 14 keynote talks, 2 hands-on sessions, and 2 panel discussions. We summarize here the ideas

[1] *Time series*, or *data series*, or *sequences* are values measured and ordered over a dimension (usually time, but could also be mass in mass spectroscopy, angle in radial chemical profiles, or position in genome sequences).

that were presented and discussed in the two workshops that took place in June and December 2016 (in Paris, France) with over 80 participants in total. The detailed program and the slides for the talks are available at the ITISA web pages.
1st edition: https://indico.in2p3.fr/event/13186/
2nd edition: https://indico.in2p3.fr/event/13934/

2. KEYNOTE TALKS

2.1 Computer Science

Prof. Anthony Bagnall (University of East Anglia) focused on Time series classification problems (TSC). He described the recent advances in time series classification, and presented a taxonomy of algorithms based on the nature of discriminatory features used to classify. Finally, he presented an experimental comparison of over 20 algorithms on 85 of the UCR-UEA datasets. The results showed that the collective of transformation-based ensembles (COTE) was significantly more accurate than all other approaches, because it could utilize features from each of the five promising representations identified in the algorithm taxonomy.

Prof. Abdullah Mueen (University of New Mexico) talked about algorithms and applications of three primitive temporal patterns, namely, motifs, shapelets, and discords. Motifs are repeating segments in seemingly random time series data; Shapelets are small segments of long time series characterizing their sources; and Discords are anomalous waveforms in long time series that do not repeat anywhere else. He discussed efficient algorithms to discover these patterns, and presented corresponding use cases. Applications included activity classification using accelerometer and brain activity data, correlated clusters in social media data, and anomaly detection in online review data.

Prof. Themis Palpanas (Paris Descartes University) presented techniques for time series indexing. He described recent efforts in designing techniques for indexing and mining massive collections of time series, and showed that the main bottleneck is the time taken to build the index. He presented the state of the art techniques that adaptively create time series indexes, allowing users to correctly answer queries before the indexing task is finished.

Prof. Anastasia Bezerianos (University Paris-Sud), Dr. Theophanis Tsandilas (Inria), and Ms. Anna Gogolou (Inria) talked about interactive visual exploration of large time series collections. They provided an overview of existing interaction and visualization techniques for time series exploration and analysis, and noted the absence of focus on their scalability to multi-terabyte time series collections. To this end, they described work directions for achieving both visual scalability (how can we visualize billions of data series) and response-time scalability (how can we get answers quickly in interactive response times), including approximate and progressive result mechanisms.

2.2 Astrophysics

Dr. Dimitrios Emmanoulopoulos (University of Southampton) talked about astrophysical time series, and in particular AGN light curves[2], whose variability allow astrophysicists to study the physical conditions around a black hole. In order to test any theoretical model though, it is crucial to attribute a precise statistical significance to any timing property. Dr. Emmanolopoulos presented a new statistical method that can produce random light curves that contain all the genuine statistical and variability properties of the observed ones, i.e., the same flux distribution (quantified by the probability density function) and same power spectral density.

Dr. Jerome Rodriguez (CEA) discussed methods for diagnosing and analyzing the fast time variability in X-ray binaries[3]. He introduced the Fourier analysis and generic techniques used in high energy astrophysics, and explained how these tools help understand the properties of fast variabilities in X-ray binaries time series.

Dr. Gabriele Ponti (Max Planck Institute) also focused on the variabilities of X-ray binaries time series. He showed that the characteristic time-scales of such variations depend linearly on the mass of the black hole, and that by studying the correlations between the emission at various energy bands (through cross spectra and lag frequency spectra), it is possible to determine delays between radiation produced by different components of the system. From this, it is possible to draw conclusions about the geometry of the regions around black holes.

Dr. Vivien Raymond (Cardiff University) focused on Gravitational Wave (GW)[4] detection using the Advanced Laser Interferometer Gravitational-wave Observatory (LIGO). He stressed that at the core of this new observational medium for GW-astronomy

[2] An *Active Galactic Nucleus (AGN)* is a compact region at the center of a galaxy that has an unusually high luminosity. A *light curve* is an astrophysical time series that measures the amount of light as a function of time.
[3] *X-ray binaries* (a.k.a. microquasars) are a class of binary stars that host the most compact objects (neutron stars and black holes), and are luminous in X-rays.
[4] *Gravitational waves* are ripples in space-time caused by violent and energetic processes in the universe (e.g., the merge of two black holes). Albert Einstein predicted the existence of gravitational waves in 1916 in his general theory of relativity, and they were first detected in 2016.

is the analysis of the time series of space-time deformations recorded by the GW detectors. He presented the time series analysis techniques currently used in signal detection, detector noise analyses, and source properties inference.

Dr. Eric Chassande-Mottin (CNRS) talked about the detection of GW in space. It is believed that the space-based LISA GW detector data stream will contain approximately 60 million simultaneous sources. He emphasized that in analyzing these time series data, there are two important goals: the first is to separate the various sources from each other, and the second is to estimate the astrophysical parameters of each source. In this task, matched filtering is the main tool of GW astronomy, a method that requires the use of accurate theoretical templates for each source type. development of sophisticated Bayesian algorithms.

2.3 Neuroscience

Dr. Katia Lehongre (ICM Institute for Brain and Spinal Cord) talked about times series analysis in neuroscience, and elaborated on the electrophysiology of epilepsy. Patients with epilepsy present abnormal brain activity, like epileptic spikes and seizures that can be recorded with electroencephalography (EEG). In order to localize the region of the brain that produces this abnormal activity, EEG from the patients is recorded continuously for 2 to 3 weeks. Usual clinical practice involves a neurologist reviewing visually the signal in order to determine the spatial localization and the temporal dynamics of the epileptic activity. As Dr. Lehongre pointed out, several studies tried to develop an automatic and reliable detection / characterization of the epileptic events in time and space, however, no fully non-supervised methods are commonly used by the neurologists, because they are not accurate enough. An efficient time series analysis could be of great interest to speed up the signal analysis, and in turn to increase the number of patients handled.

Prof. Uri Hasson (University of Trento) discussed time series in relation to the brain. He began with a brief overview of the sorts of time series that modern cognitive neuroscience can obtain from human participants and the principles of the instrumentation used to obtain those. The second part of the talk focused on approaches for analyzing these times series. These include quantification of correlations between different brain regions and network partitioning strategies. It was noted that more recent work is focusing on fast, non-oscillatory signatures in brain dynamics that also contain important information. These signatures are either driven by an input or internally generated. Several methods based on the analysis of peaks and pits in neural time series were discussed, as well as methods for decomposing spatiotemporal data into series of micro-states and motifs. The last part summarized these technologies from the perspective of temporal search engines, highlighting the importance of approximate searches on multivariate time series, and in the context of real-time analysis.

2.4 Engineering and Electricity Networks

Dr. Dohy Hong (Safran) shared his experience on multivariate time series analysis in aeronautics, and aircraft engines in particular. He pointed out that one of the main future challenges in aeronautics is the use of available data in order to enable the optimal management of maintenance processes (e.g., a per engine-based individual management strategy), and later its integration in the design process. The overall data processing along the engine cycle includes several technical challenges: weak signal detection in continuous multivariate usage time series (hidden layer/rules learning/extraction), management of heterogeneous data (maintenance repair and operation data, test bench or inspection data, configuration data, etc.), integrating and consolidating the existing expert knowledge (from design model to residual life time estimate practice), and others. All these challenges have to be addressed under the constraint of certifiability, which implies interpretability and robustness.

Dr. Georges Hebrail (EDF) made the case for privacy-preserving use of individual smart meter data for customer services. The advent of on-body/at-home sensors connected to personal devices leads to the generation of fine grain highly sensitive personal data at an unprecedented rate. However, despite the promises of large scale analytics there are obvious privacy concerns that prevent individuals to share their personal data. Dr. Hebrail presented Chiaroscuro, a solution for clustering personal time series, with strong privacy guarantees. The execution sequence produced by Chiaroscuro is massively distributed on personal devices, coping with arbitrary connections/disconnections. Chiaroscuro builds on a novel data structure, which allows the participating devices to collaborate privately by combining encryption with differential privacy.

2.5 Music

Prof. Philippe Esling (IRCAM) talked about musical time series. Music inherently conveys several open and interesting scientific questions, which all embed the notion of time. Specifically, musical orchestration is the subtle art of writing musical pieces

for orchestra, by combining the spectral properties specific to each instrument in order to achieve a particular sonic goal. Prof. Esling described novel learning and mining algorithms on multivariate time series that can cope with the various time scales that are inherent in musical perception, and can be used for orchestration. His current research is focused on automatic inference through deep representational learning to allow the automatic deciphering of these dimensions in order to provide optimal features for orchestration, by targeting correlations existing in the work of notorious composers.

3. HANDS-ON SESSIONS

The workshop also offered two hands-on sessions.

The first hands-on session was organized by Dr. Vivien Raymond, and walked participants through the process of analyzing real time series signals collected at LIGO, and visualize a GW that was buried in the signal. This session aimed at teaching participants all the preprocessing steps necessary for cleaning the time series, and for amplifying the true signal, i.e., the GW. This process involved filtering and noise removal and downsampling steps[5], which had to be performed either in the time- or frequency-domain. The main take-away message of this session was that the time series processing workflow that analysts apply (often times) requires many preprocessing and data transformation steps.

The second hands-on session was led by Ms. Anna Gogolou, and dealt with the challenges in interactive visual exploration of large time series. The participants were asked to reply to a questionnaire that was divided in three parts: background information, scenarios, and detailed examples of their questions and problems at hand. Participants came mainly from two different domains: neuroscience and astrophysics. Analyzing their answers, led to the conclusion that their goal on multi-dimensional time series data is to find: similar patterns, abnormal patterns, time length of events, specific times of variability in data (periodicity), and correlation. Moreover, some are interested in working with quick, but rough results (e.g., approximations or incomplete answers), while they wait for the complete and exact answer.

4. DISCUSSION SESSIONS

Finally, we report on the discussion sessions of the workshop, which greatly helped in putting all previous information in perspective, and in identifying the research directions that are useful and promising. Below, we summarize the main points of the discussion, and relevant open research problems.

[5]`https://www.gw-openscience.org/tutorials/`

(1) Preprocessing: In most cases, time series must be preprocessed before being analyzed: this involves selecting the series and intervals of interest, and applying techniques from signal processing (e.g., band filters, denoising), several of which operate in the frequency domain. Thus, time series management systems [12] and analysis workflows should allow analysts to easily extract subsets of interest, and embed their data in different spaces (e.g., time, frequency), suitable for the various analysis techniques.
(2) Analysis Operations: The analysis task itself encompasses several different operations, including similarity search, correlation, clustering, classification, anomaly (or discord) detection, motif (or frequent patterns) discovery, and causal modeling[6]. Similarity search is a key operation that is expensive per se, and if performed fast then it can help speedup several of the other analysis operations, as well. We note that in some cases, the analysis operations need to be performed in a way that takes into account spatial information[7], pointing to the need for the development of corresponding query languages and index structures.
(3) Versatility: Even though there exists a sizable number of studies on time series analysis techniques in the literature (and some of them have found their way into real systems), these techniques are usually not versatile enough for use in the real world. Real applications require scalable techniques that can serve ad-hoc queries and analysis workflows, have the ability to select and operate on sets of sequences selected using complex conditions[8], operate on both entire sequences and subsequences, support the analysis of variable-length subsequences[9], treat value uncertainty[10] [4] as a first class citizen and be able to carry confidence and significance values along the analysis workflow, as well as allow for privacy-preserving analytics.
(4) Interaction with Users: Despite the significant effort of the visualization community in this area [2], most of the available systems are far from scaling to the sizes of time series collections that are used in practice. Recent advances in time series indexing can be of help here, though, novel approaches are required in order to really address the current and future needs. Interactive visualizations are important for users, and when the datasets

[6]For example, Granger causality.
[7]Neuroscientists are interested in correlations among signals recorded by sensors that are spatially close.
[8]Based on metadata, time intervals, value thresholds.
[9]Consider for example that most of the current time series indexes only support fixed-length queries.
[10]In several cases, this uncertainty is inherent in the measurement instrument.

grow very large some of the most promising ways to achieve interactive response times is through the use of fast approximate and/or progressive answers [27], with the support of appropriate visualizations.

[**Summary**] Overall, we observed a slight disconnect between the needs of scientists and practitioners that process and analyze time series, and Computer Science (CS) researchers that work on time series. The problems that CS researchers have been studying are for the most part simplified, clean, and sanitized versions of the real problems and analysis workflows that practitioners have to address in the real world. Despite the remarkable progress in this area by the CS community during the recent years [6, 8, 12, 14, 18, 22], there are still many challenging open problems.

Efficiently supporting similarity search [3, 16, 17, 23,24,28] is still challenging for large data series collections [6,7]. Only very recently attention has been given to solutions that can support variable-length queries [19, 20], and there are still a lot to be done in terms of supporting uncertain series [5]. Scalable visualization solutions are direly needed, especially in support of progressive analytics [10, 27]. At the same time, even some basic problems, such as the interplay between visual perception and similarity measures [9], deserves to be studied in more detail. Evidently, in order to be used in practice, all the above components should be combined in general, easy-to-use by non-experts, time series management systems [6, 12], a task that is by itself a challenge.

5. WORKSHOP CONCLUSIONS

Even though time series are a very common data type, no available system can inherently accommodate and support the dataset sizes and complex analytics required by users. Our discussions showed that applications across different domains share common requirements: fulfilling them is a challenging goal, involving many interesting research problems.

[**Acknowledgements**] The workshops were supported by the CNRS Mastodons TimeClean project.

References

[1] R. Agrawal, C. Faloutsos, and A. N. Swami. Efficient similarity search in sequence databases. In *FODO*, 1993.

[2] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Springer, 2011.

[3] 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.

[4] M. Dallachiesa, B. Nushi, K. Mirylenka, and T. Palpanas. Uncertain time-series similarity: Return to the basics. *PVLDB*, 5(11), 2012.

[5] M. Dallachiesa, T. Palpanas, and I. F. Ilyas. Top-k nearest neighbor search in uncertain data series. *PVLDB*, 8(1), 2014.

[6] 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.

[7] K. Echihabi, K. Zoumpatianos, T. Palpanas, and H. Benbrahim. Return of the Lernaean Hydra: Experimental Evaluation of Data Series Approximate Similarity Search. *PVLDB*, 2019.

[8] C. Faloutsos, J. Gasthaus, T. Januschowski, and Y. Wang. Forecasting big time series: Old and new. *PVLDB*, 11(12), 2018.

[9] A. Gogolou, T. Tsandilas, T. Palpanas, and A. Bezerianos. Comparing similarity perception in time series visualizations. *TVCG*, 25(1), 2019.

[10] A. Gogolou, T. Tsandilas, T. Palpanas, and A. Bezerianos. Progressive similarity search on time series data. In *EDBT BigVis Workshop*, 2019.

[11] P. Huijse, P. A. Estévez, P. Protopapas, J. C. Principe, and P. Zegers. Computational intelligence challenges and applications on large-scale astronomical time series databases. *IEEE Comp. Int. Mag.*, 9(3):27–39, 2014.

[12] S. K. Jensen, T. B. Pedersen, and C. Thomsen. Time series management systems: A survey. *TKDE*, 29(11), 2017.

[13] K. Kashino, G. Smith, and H. Murase. Time-series active search for quick retrieval of audio and video. In *ICASSP*, 1999.

[14] E. J. Keogh. Indexing and mining time series data. In *Encyclopedia of GIS.*, pages 933–939. 2017.

[15] E. J. Keogh, K. Chakrabarti, S. Mehrotra, and M. J. Pazzani. Locally adaptive dimensionality reduction for indexing large time series databases. In *SIGMOD*, 2001.

[16] H. Kondylakis, N. Dayan, K. Zoumpatianos, and T. Palpanas. Coconut Palm: Static and Streaming Data Series Exploration Now in your Palm. In *SIGMOD*, 2019.

[17] H. Kondylakis, N. Dayan, K. Zoumpatianos, and T. Palpanas. Coconut: Sortable Summarizations for Scalable Indexes over Static and Streaming Data Series. *VLDBJ*, accepted for publication, 2019.

[18] J. Large, P. Southam, and A. J. Bagnall. Can automated smoothing significantly improve benchmark time series classification algorithms? *CoRR*, abs/1811.00894, 2018.

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

[20] M. Linardi, Y. Zhu, T. Palpanas, and E. J. Keogh. Matrix profile X: VALMOD - scalable discovery of variable-length motifs in data series. In *SIGMOD*, 2018.

[21] T. Palpanas. Data series management: The road to big sequence analytics. *SIGMOD Record*, 44(2), 2015.

[22] T. Palpanas. Big sequence management: A glimpse of the past, the present, and the future. In *SOFSEM*, 2016.

[23] B. Peng, P. Fatourou, and T. Palpanas. Paris: The next destination for fast data series indexing and query answering. In *IEEE BigData*, 2018.

[24] B. Peng, P. Fatourou, and T. Palpanas. MESSI: In-Memory Data Series Indexing. In *ICDE*, 2020.

[25] U. Raza, A. Camerra, A. L. Murphy, T. Palpanas, and G. P. Picco. Practical data prediction for real-world wireless sensor networks. *TKDE 27(8)*, 2015.

[26] D. Shasha. Tuning time series queries in finance: Case studies and recommendations. *DEBull*, 22(2), 1999.

[27] 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.

[28] D.-E. Yagoubi, R. Akbarinia, F. Masseglia, and T. Palpanas. Massively distributed time series indexing and querying. *TKDE (to appear)*, 2019.

[29] L. Ye and E. J. Keogh. Time series shapelets: a new primitive for data mining. In *KDD*, 2009.

[30] K. Zoumpatianos, S. Idreos, and T. Palpanas. ADS: the adaptive data series index. *VLDBJ*, 25(6), 2016.

[31] K. Zoumpatianos and T. Palpanas. Data series management: Fulfilling the need for big sequence analytics. In *ICDE*, 2018.