Themis Palpanas Speaks Out on Work, Collaborations, and Enjoying Life Opportunities

H. V. Jagadish and Vanessa Braganholo

Themis Palpanas

https://helios2.mi.parisdescartes.fr/~themisp/

Welcome to ACM SIGMOD Record's series of interviews with distinguished members of the database community. I'm H. V. Jagadish, Professor of Computer Science at the University of Michigan. Today I have the honor of interviewing Themis Palpanas, who is a Distinguished Professor of Computer Science at Université Paris Cité. He is a Senior Fellow of the French University Institute (IUF), the Head of the Computer Science Department at the Université Paris Cité, and the Director of the Data Intelligence Institute of Paris (diiP). Themis, welcome!

You have done a lot in the database field. We are eager to learn a little bit about your work and your thoughts about the field. One topic that immediately comes to mind is time series analytics. You've been working in that for over two decades, so can you talk a little bit about the advances for data management in that field?

Thank you very much for the invitation, Jag. It's an honor to be part of this series.

For the last several years, time series analytics[1] has been the core focus of the work in my lab, and with my collaborators. What is interesting about time series is that it includes several different challenging data management problems. So this is what got me really excited since the first time that I got into this area, and I'm still excited to work on this now.

It's not an easy data management problem for two main reasons. One is that we're talking about a special data type that is very high dimensional. You can think of a time series as a long sequence of real values. This sequence can be thought of as a vector, right? So we are talking about a high dimensional vector, and it does not matter if we're talking about a large collection of small vectors or a single, very long series or infinite series. In either case, the patterns of interest (the patterns that we want to identify and analyze) are in the order of several hundreds to several thousands of points. And this basically defines the dimensionality of the space in which we need to work. So, we have these high dimensional spaces of hundreds to thousands of dimensions – this is the first challenge.

The second challenge is that the datasets that we want to work with are often very large: they are in the order of terabytes, or even petabytes. There are plenty of examples of these across all disciplines and domains. To give you an idea, I can mention astrophysics. You may have heard about the gravitational waves that were recently detected for the first time. A gravitational wave is nothing else but a time series. What is even more interesting is that the machinery that the physicists have set up to be able to detect these series is so extensive and so complex that it needs to monitor itself to make sure that everything works correctly. This machinery involves in the order of 10,000 additional streaming series produced by sensors, which monitor the operational health of the machine that detects gravitational waves. Obviously, all these series need to be analyzed as fast as possible. In some particular cases, we are interested in analyzing these signals in near real-time, because we may end up detecting some interesting signal that would allow us to then turn on another kind of telescope, for example, gamma-ray telescopes, towards the source that we have identified. There is a window of a few minutes when this could be done. So, there is a lot of interest in this community in having very accurate and also very scalable ways of analyzing all these time series.

In time series analysis, similarity search, clustering, classification, frequent patterns, and anomaly detection are some of the very interesting and challenging problems that the community is working on. Similarity search is very often employed in these other kinds of analysis as well. For example, k-NN classification is based on similarity search.

If we take a look at similarity search (this has also been the main focus of our own work), there are several different subproblems[2]. For example, what happens when you are interested in different kinds of distances? Some applications may use Euclidean distance. Some other applications may use some elastic distance measure that allows you to match interesting patterns, even if they are not aligned in time (e.g., Dynamic Time Warping (DTW)). Having picked our distance, there are similarity search flavors depending on the length of the (data and query) series. We may have a large collection of small series to analyze, or we may have a single long series, where we need to look at all its subsequences. We need different solutions for each of these cases. Do we want to do whole matching (match the entire query against some candidates), or do we want to do subsequence matching (match part of the query or part of the candidate)?

We also have different kinds of query-answering solutions. We can have exact queries, where we always return the exact answer with probability one, but we also have approximate queries with several different flavors[3]. They range from approximate queries with deterministic guarantees –with probability one, return answers within an error ε of the exact answer–, or we can have approximate queries with probabilistic guarantees, or even approximate queries with no guarantees whatsoever. This last type of similarity

[1] Themis Palpanas, Volker Beckmann. Report on the First and Second Interdisciplinary Time Series Analysis Workshop (ITISA). ACM SIGMOD Record 48(3), 2019.

[2] 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).

[3] 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).

queries, approximate with no guarantees (ng-approximate), may initially sound strange, but is widely used in practice: the answers that they give back are, most of the time, very close to the exact answers, and they are much faster than the other flavors of similarity search. So, in order to favor speed, several applications may drop the quality guarantees.

Interestingly, as the hardware started changing, we also had to look at this dimension as well: study these problems related to different hardware configurations. What happens when the dataset is in-memory? What happens if the dataset cannot fit in main memory? How do we parallelize when we are in a single node? How can we take advantage of GPUs? What happens if we go to a distributed setting?

It is true that different deep learning solutions have been applied to most of these problems, especially on the traditional mining learning tasks (clustering, classification, forecasting, anomaly detection). But it seems that we are not yet at the point where we should throw away the traditional solutions

There has been lots of work on all these different problems in the last twenty years, and these problems have been the main focus of the work in my group. There now exist algorithms that are pretty efficient for all these situations, and we developed several of the state-of-the-art solutions for the entire spectrum of these problems[4].

In the past 20 years, one particular type of similarity search, exact search, has been sped up by 2-3 orders of magnitude. What is most interesting is that the progress that we have made for this problem was all due to ideas coming from data management: how to best organize and then access the data.

Note that several different sub-communities are related to this problem, including data management, information retrieval, time series, and machine learning. Personally, I started looking at this problem by studying the literature in the time series community. Though, remember that, conceptually, time series are vectors. As such, all the work that we have done in data management in the area of multidimensional points (e.g., R-trees, k-d-trees, X-trees, M-trees, LSH) is relevant. Recently, another community working on this problem proposed a graph representation, the k-NN Graphs, and corresponding solutions.

Just a few years ago, my group conducted the first study that looked at the solutions coming from all these different communities[2,3]. What was really surprising for me was to actually see that the techniques that we have been developing for time series were working extremely well for general high-dimensional vectors, as well. It is now very interesting that we are at a point where we can close the loop, study the solutions from all these communities together, compare them, and learn from one another. I find this very exciting, and we already have high dimensional vector indexes using such cross-pollinated ideas with very promising results[5,6,7].

This is a crucial observation going forward, because general high-dimensional vectors are now used widely for indexing and searching large collections of deep embeddings. We can now embed any complex object (e.g., video or image) into a high dimensional vector, and then we can analyze these objects in the embedded space, since it is much easier doing similarity search of vectors instead of the original videos. Then suddenly, this kind of complex analytics with any kind of object becomes easier and faster, because they are now based on high-dimensional vectors. All the work that we have been doing is very relevant to this case as well.

There are two Special Issues in the IEEE Data Engineering Bulletin, in September 2023[8] and September 2024[9]. Whoever is working on this field should read this collection of papers. They talk about several of these different solutions and how they relate to one another. So, I think that is a very exciting area to work on, with many real and challenging applications.

[4] Themis Palpanas. Evolution of a Data Series Index - The iSAX Family of Data Series Indexes. Communications in Computer and Information Science (CCIS) 1197, 2020.

[5] Ilias Azizi, Karima Echihabi, Themis Palpanas: Elpis: Graph-Based Similarity Search for Scalable Data Science. Proc. VLDB Endow. 16(6): 1548-1559 (2023).

[6] Jiuqi Wei, Botao Peng, Xiaodong Lee, Themis Palpanas: DET-LSH: A Locality-Sensitive Hashing Scheme with Dynamic Encoding Tree for Approximate Nearest Neighbor Search. Proc. VLDB Endow. 17(9): 2241-2254 (2024).

[7] Qitong Wang, Ioana Ileana, Themis Palpanas: LeaFi: Data Series Indexes on Steroids with Learned Filters. Proc. ACM Manag. Data 3(1): 51:1-51:27 (2025).

[8] http://sites.computer.org/debull/A23sept/issue1.htm

[9] http://sites.computer.org/debull/A24sept/issue1.htm

You mentioned a number of technologies that you would bring to bear from many different areas, but notably, you didn't mention anything about AI, which seems to be so much in the news these days. How do you feel about neural networks and LLMs? So, for example, could you use LLMs to analyze, say, news and correlate them with the stock market values or political events, you know, things like this?

Yes, definitely. All these kinds of solutions are now extremely popular. It is true that different deep learning solutions have been applied to most of these problems, especially on the traditional mining learning tasks (clustering, classification, forecasting, anomaly detection). But it seems that we are not yet at the point where we should throw away the traditional solutions. In the last couple of years, we have started seeing different studies that compare all these methods.

I think that the overall conclusion is that there is no single best solution across a wide range of different data sets. But even more importantly, it is not at all certain that deep learning is doing better than traditional methods. Mind you that deep learning oftentimes needs training that some traditional methods do not need; or it needs more training than traditional methods. So, I think that there is still no final verdict on this. However, I'm not against machine learning and deep learning. All these techniques come with a certain promise – they can adapt to different kinds of data (with a demonstrated positive impact on high dimensional vector similarity search[10] and anomaly detection[11]). They can also learn by themselves, and lead, for example, to anomaly detection solutions with no explicit user-specified algorithm on how to define or find anomalies.

Another important point is that deep learning methods can naturally handle multivariate data series. Usually, when we talk about time series, we have in mind some time series where each point in this time series is a scalar; it is a single real value. But each one of these points can also be a vector of values. We call these series *multivariate*. For example, a sensor that produces temperature, humidity, and vibration. Deep learning, in particular, is very good at handling multivariate series. This is important because going multivariate with traditional techniques, for many of them (if not all of them), means that the complexity explodes, either time complexity or space complexity – usually both. The community has started looking at how we can integrate these kinds of ideas in this context, and my group, as well[12]. Once again, I think that this is a very promising research direction: it gives us the opportunity to inherently process multivariate datasets, and to become more data-adaptive, which translates to increased efficiency.

You just said something about the interaction between time series and data management and thinking about it as two separate things. So, I'd like to understand how you feel about the DB community and the kind of work that you do and others do on time series data analysis. Is it a good relationship? Would you like to change things?

To clarify my point, I do not consider time series separate from data management. There are several commercial data management products nowadays focusing on time series management. These are systems that cater to the IoT kind of applications, or to operational health monitoring. Though, in the context of these systems, there is still lots of research work to be done. There is work on building declarative interfaces, as well as on the backend of these systems in terms of optimizing the operations they need to perform and their execution. There are no sophisticated, optimized solutions for similarity search; the same for other kinds of more complex analytics, including clustering and classification. Having said that, it is also true that there are other communities that are relevant here, such as machine learning and data mining, but this has been true in the past as well for the mining and analysis of structured data.

Another point here is that if you observe the different data management conferences, there is usually no explicit mention to time series. In the list of topics, time series papers are treated as papers under the "temporal databases" category. However, this is not exactly what time series are. There are differences between temporal databases and all the methods we use for time series analysis. This year, VLDB explicitly mentions both "time series" and "high-dimensional vectors" in the list of topics, which I feel is very important.

Besides time series, you have done a lot of work on entity recognition and data integration. Would you like to talk a little bit about that area?

[10] Qitong Wang, Themis Palpanas: SEAnet: A Deep Learning Architecture for Data Series Similarity Search. IEEE Trans. Knowl. Data Eng. 35(12): 12972-12986 (2023).

[11] Emmanouil Sylligardos, Paul Boniol, John Paparrizos, Panos E. Trahanias, Themis Palpanas: Choose Wisely: An Extensive Evaluation of Model Selection for Anomaly Detection in Time Series. Proc. VLDB Endow. 16(11): 3418-3432 (2023).

[12] Paul Boniol, Mohammed Meftah, Emmanuel Remy, Themis Palpanas: dCAM: Dimension-wise Class Activation Map for Explaining Multivariate Data Series Classification. SIGMOD Conference 2022: 1175-1189.

This is another topic on which I have been working for more than 10 years, and I have to acknowledge my collaborator Dr George Papadakis, who has been the main driving force behind this work. In this area, our focus has been on scalability in entity resolution[13]. Just to give some context, in entity resolution, we need to identify whether or not two entities refer to the same real-world object. In general, when we have a collection of entities, we need to perform a quadratic number of comparisons (all to all), to figure out which of these entities are the same. One way of scaling this problem is by performing blocking, that is, grouping similar entities together so that when we want to compare entities, we only compare the entities that belong to the same block.

> ***[T]here are two ingredients that are important [to foster collaborations]. You need quite a bit of patience, and you also need some luck. I guess I have had both!***

In this context of blocking, we have developed solutions that are scalable and domain-agnostic. One particular method that we proposed is Meta-blocking[14], which takes as input a set of blocks, and transforms it into a new set of blocks that drastically reduces the number of entity comparisons, while attaining essentially the same recall. Meta-blocking is based on the idea of representing a blocking solution as a graph (where entities are represented as nodes, and edges connect entities that share at least one common block in the original blocking solution), and then manipulating this graph to eliminate superfluous comparisons.

This technique has been proven extremely efficient for different kinds of data, including unstructured data, where there are no specific attributes for each entity. It has been used in online settings for progressive entity resolution[15], and has also been extended to supervised meta-blocking[16], where you have a machine-learning technique that tells you how to prune this meta-blocking graph to end up with the final set of blocks.

The above methods, as well as the related work and state-of-the-art techniques are included in a book that describes all these solutions: The Four Generations of Entity Resolution[17].

We've been talking about the fact that time series is interdisciplinary, and you've had different areas in which you have worked. You mentioned that the work that you were doing on entity resolution was collaborative with another person that you gave credit to. It appears that you have a lot of collaborations. You are able to initiate new ones very easily, given how readily you are giving credit to a collaborator. Do you care to tell us how you think about collaborations?

I should start by saying that there are two ingredients that are important here. You need quite a bit of patience, and you also need some luck. I guess I have had both!

My starting point is that not all collaborations will be fruitful. Nevertheless, I enjoy getting in this kind of collaborative work and trying to see where it will lead me. To give you one example, I was out with some friends for a cup of coffee, when an acquaintance of one of my friends arrived. This person was a physicist working on the mass spectrometry of apples, and mentioned that he had lots of mass spectra of apples. Mass spectra data are essentially data series, where the *x* value is not time – it is mass. This is luck: there is this guy that has a collection of this kind of data series and he wants to perform similarity search, and just out of the blue, we started this discussion, which initially led to a small prototype for them to use. This allowed us to identify some issues with the solutions in the literature, and that got the ball rolling, leading to a 15-year research effort, with 12 MSc and 8 PhD theses, on the problem of data series similarity search! I definitely believe that talking to people from other disciplines is extremely useful. That's where patience comes into play, because when you start this kind of discussions, there is always a gap in the vocabulary, in the way that

[13] George Papadakis, Georgios M. Mandilaras, Luca Gagliardelli, Giovanni Simonini, Emmanouil Thanos, George Giannakopoulos, Sonia Bergamaschi, Themis Palpanas, Manolis Koubarakis: Three-dimensional Entity Resolution with JedAI. Inf. Syst. 93: 101565 (2020).

[14] George Papadakis, Georgia Koutrika, Themis Palpanas, Wolfgang Nejdl: Meta-Blocking: Taking Entity Resolutionto the Next Level. IEEE Trans. Knowl. Data Eng. 26(8): 1946-1960 (2014).

[15] Giovanni Simonini, George Papadakis, Themis Palpanas, Sonia Bergamaschi: Schema-Agnostic Progressive Entity Resolution. IEEE Trans. Knowl. Data Eng. 31(6): 1208-1221 (2019).

[16] Luca Gagliardelli, George Papadakis, Giovanni Simonini, Sonia Bergamaschi, Themis Palpanas: GSM: A generalized approach to Supervised Meta-blocking for scalable entity resolution. Inf. Syst. 120: 102307 (2024).

[17] George Papadakis, Ekaterini Ioannou, Emanouil Thanos, Themis Palpanas: The Four Generations of Entity Resolution. Synthesis Lectures on Data Management, Morgan & Claypool Publishers 2021, ISBN 978-3-031-00750-7, pp. 1-170

people understand, or value things. But given enough time, if you reach the point where you can actually understand the real problems the other discipline is working on and how you can help them, then you can build trust, and then it can become a very fruitful collaboration. Such interdisciplinary collaborations give rise to new questions for the data management perspective as well.

One of the things that also comes out from all of this is both the breadth in terms of the range of knowledge that you have and that you seek as a goal in itself because you enjoy it. Do you think that's a fair characterization in terms of you really value breadth?

I will take this as a compliment! I cannot really comment on that. What I can say is that I really enjoy working on different problems. For example, I am not only interested in core data management problems, but also in everything that has to do with analytics, including mining and machine learning. By bringing them together, you end up with very exciting research problems, as well as collaborations.

Turning to your own institution, you've set up a major research institute and built up a very successful research group starting from scratch a few years ago. Can you talk about your journey there and things that led to your great success in that direction?

You are referring to the Data Intelligence Institute of Paris (*diiP*). What is interesting is that diiP started its operations right before the Covid Pandemic, which was a little bit challenging. It did not let us have the kind of face-to-face interactions that we were hoping for. Nevertheless, it survived, and it has now grown into an institute that's well-regarded. The goal of diiP is to support interdisciplinary projects that are related to data science and data intelligence. What we want to do is to provide researchers, who are not data intelligence experts, with the necessary expertise and the means to achieve results when analyzing their data. In the past three and a half years, we have supported more than sixty interdisciplinary projects, and we have organized several workshops, seminars, and hands-on sessions.

Talking about organizing things, you have organized so many things. You're chairing conferences and being very active in the community. How do you manage to juggle all of these things in terms of how do you manage time so successfully?

Well, I'm not sure that I actually manage my time very successfully: I may not be the right person to talk about work-life balance! Work has been taking quite a bit of time in my life. This has been especially true in the last years, after I joined this position in Paris. It turned out that the opportunities here were too exciting to pass on. I tried to get the most out of them, and this has led me to really overwork myself. The answer to your question is that I just put too many hours into my work. At the same time, I have been blessed with some excellent students and collaborators.

As I was preparing for this interview, your students had wonderful things to say about you. Of course, most students appreciate their advisor, but I thought it was more than that. So, I want to say that you are very much appreciated by your students. Do you have any comments or any thoughts about why that might be the case?

I am very happy to hear that! It is true that I always try to be close to my students, in the sense that we do not only meet when we have to discuss work. In general, I am trying to foster a sense of community inside our group. We often organize outings: sometimes we go for lunch in the gardens of the Louvre, which is extremely nice; we also do different activities, for example, play group games together.

Another point I wanted to make is that, of course, not all students are created equal. I feel that my role is to try to push each student a few steps further than the point they thought they could reach. I believe that this resonates with them: trying to get the best out of each student, and trying to make each one of them evolve during this journey towards their PhD.

Moving on to bigger life issues. You have a unique perspective, I think, amongst database researchers of having lived and worked for substantial periods in multiple countries. So I wanted to have a little bit of benefit of insight from you in terms of how do you compare the various places that you have been to and how do you choose to move?

I did my undergraduate studies in Athens, Greece, and my graduate studies in Toronto, Canada. I then moved to the Los Angeles area, USA (University of California, Riverside). I worked for a couple of years at IBM Watson Research Center in New York, USA. I then moved to the University of Trento, in Italy, and subsequently to the Université Paris Cité, France. So, your observation is correct, but honestly, it is not easy to move around. It is not just about changing workplaces. It is about moving your entire life: you have to restart your life in every new place. Moreover, as you may imagine, this becomes harder as you grow older. I did most of these moves when I was much younger, and there was lots of excitement involved in all this. I should say that I have no second thoughts about having moved

around all these places. I enjoyed a lot working in these different places, as well as living in and experiencing these places, peoples, and cultures. Each one of these places offered very different options. The key in enjoying such a journey is to make use of the options offered to you. I never moved to a new place expecting to live the kind of life that I was living earlier. I always try to adapt to the new ways of life, to the new opportunities, and this process of adapting is very enriching, because you end up discovering new ways of finding and appreciating the beauty in life, as well as the beauty of life.

In this context, if you look at the places you have lived, you've mentioned Athens, Toronto, New York, LA, and now Paris. How does Trento fit in this series? You know, with all of these big cities, Trento is a very small place.

Right, Trento is the outlier here, and (given my work on anomaly detection) there had to be one! Trento is an interesting story. It happened serendipitously. It was a point where I was looking to go back from North America to Greece. I was in the process of exploring my options at the Greek universities, and preparing applications for those. In the middle of this process, when I was mentally prepared to leave from North America, the Trento opportunity popped up. I decided to visit them, without knowing what to expect, and I was happily surprised by the research environment and mentality. Workwise, it was an environment that I appreciated. It was very particular for the Italian context, and it definitely helped me take the first steps in my academic career, and establish myself. Together with Prof. Yannis Velegrakis, we set up the dbTrento group; it was a very exciting period of time. Life-wise, you can imagine that the Trento experience would be extremely different for someone who moved from New York. Trento is a city of a hundred thousand people. While New York is all about going out in the city and meeting up with all sorts of different people from all different corners of the world, Trento is all about outdoor activities, and I did enjoy those a lot: going to the mountains, both during summertime and wintertime, doing snowboarding, hiking to (and swimming in) the lakes. The Trento area is an extremely beautiful part of the world, and very dear to my heart!

And I believe that Trento has memories of you in the form of your photograph collection in the CS department in the university, and, in some other places, I believe in a nursing home and so on. So, could you say a little bit about your photography hobby and how it started, and are you still continuing?

> ***I feel that my role is to try to push each student a few steps further than the point they thought they could reach.***

Yes, photography is a very dear hobby. It started when I was in the United States, when I got my first digital camera. We could maybe say that I am an amateur photographer, nothing more than that. While in Trento, I tried to pursue this hobby further, so I got involved in some group exhibitions, and also organized some personal exhibitions. Like you mentioned, two of them are permanent. There is a small exhibition at the Department of Computer Science at the University of Trento, *Journey Towards Knowledge*[18], dedicated to PhD students. That was a new building with lots of white walls, which I volunteered to decorate. There is also another permanent exhibition, *Window to the World*[19], at a nursing home near Trento. That was a project for bringing life to the walls of a newly constructed section for this nursing home.

If you want to take the next step with any hobby, with photography, in this case, you need to invest a considerable amount of time. It is not only about taking the pictures: you need to process them; you need to build up your presence as a photographer in order to be able to talk to other people and showcase your work, to participate in exhibitions or organize exhibitions; you need to have a corresponding CV and website. All this really takes lots of time. Unfortunately, during the last years, I have not been able to dedicate to photography as much time as I would have liked.

Besides photography, you mentioned snowboarding in Trento, and I hear that you're also very good dancer, like with Latin dancing and so on. Is there things you'd like to tell us about some of your other hobbies?

This is part of our discussion on how to make the most out of the opportunities that a place offers you. The first time that I tried snowboarding was in Toronto; well, not in Toronto, but in the mountains of Quebec. As a graduate student I did not have many opportunities to practice snowboarding. I did more of that when I was in the United States, and snowboarding became one of my prime wintertime activities when I was in Trento: the

[18] https://tinyurl.com/DisiCollection

[19] https://tinyurl.com/ClesCollection

Dolomites, the mountains that surround Trento, are very beautiful. I still try to go back there for snowboarding every year. This is something that I enjoy a lot.

Latin dancing came about in Toronto, where I had several friends, fellow students, from South America. We were going out to places with latin music, and I very much liked the vibe. So, I picked it up in the same way that I also tried to pick up Spanish. Then, when I moved to Italy, I had to learn Italian; now in France, French is necessary. All this variety contributes for a rich life experience, which I enjoy tremendously.

And I really appreciate you for that. So, you're supposed to be foodie, is what I've heard. And you're in a city where definitely people talk about food, right now. Do you have secrets from Paris that you want to share?

Well, I don't think I have any secrets. There are some places that I enjoy going to, and sometimes, I make an effort to go to these particular places for different reasons: the kind of ambiance, or some particular types of food that they are serving. It is really interesting to experience the French cuisine in its different flavors, and I definitely enjoy contemporary French cuisine. I should add that I also like a lot the Italian cuisine: I admire the miraculous way in which they use very simple ingredients, they put them together with very little processing, and the outcome is outstanding. I really appreciate the Italian cuisine for that.

So, Themis, thank you so much for speaking out with us today.

Thank you very much!.