A Summary of ICDE 2022 Research Session Panels

Zhifeng Bao, Panagiotis Bouros, Reynold Cheng, Byron Choi, Anton Dignös, Wei Ding,Yixiang Fang, Boyang Han, Jilin Hu, Arijit Khan, Wenqing Lin, Xuemin Lin, Cheng Long, Nikos Mamoulis, Jian Pei, Matthias Renz, Shashi Shekhar, Jieming Shi, Eleni Tzirita Zacharatou, Sibo Wang, Xiao Wang, Xue Wang, Raymond Chi-Wing Wong, Da Yan, Xifeng Yan, Bin Yang, Dezhong Yao,Ce Zhang, Peilin Zhao, Rong Zhu

Abstract

In the 38th IEEE International Conference on Data Engineering (ICDE), 2022, panel discussions were introduced after paper presentations to facilitate in-depth exploration of research topics and encourage participation. These discussions, enriched by diverse perspectives from experts and active audience involvement, provided fresh insights and a broader understanding of each topic. The introduction of panel discussions exceeded expectations, attracting a larger number of participants to the virtual sessions.

This article summarizes the virtual panels held during ICDE'22, focusing on sessions such as Data Mining and Knowledge Discovery, Federated Learning, Graph Data Management, Graph Neural Networks, Spatial and Temporal Data Management, and Spatial and Temporal Data Mining. By showcasing the success of panel discussions in generating inspiring discussions and promoting participation, this article aims to benefit the data engineering community, providing a valuable resource for researchers and suggesting a compelling format of holding research sessions for future conferences.

1 Introduction

The 38th IEEE International Conference on Data engineering (ICDE) was hosted in Kuala Lumpur, Malaysia from May 9to May 12 virtually in 2022. To provide a platform for experts and thought leaders from different backgrounds to explore different research topics thoroughly and promote participation, panel discussion was conducted for each session after paper presentation in ICDE'22. The sharing of diverse perspectives from experts and active audience participation enriched the discussion, providing a broader understanding of each topic and offering fresh insights that are beneficial for future research directions. The introduction of panel discussions for research paper presentation sessions also attracted more participants than expected for each virtual session.

Since the panel discussion introduced in ICDE'22 is quite successful in both generating inspiring discussions and promoting participation of research sessions, it would benefit the data engineering community by summarising the discussions of the panels of ICDE'22, and this format of conducting research sessions would be of interest to future conferences. This article summarised the virtual panels held for some interesting sessions in ICDE'22. The organizations of this paper is as follow: Section 2 and 3 summarises the Data Mining and Knowledge Discovery session while Section 4 talks about the Federated Learning session. Graph Data Management session and Graph Neural Network session are discussed in Section 5 and Section 6. Section 7 and Section 8 summarise the Spatial Temporal Data Management session and Spatial and Temporal Data Mining session.

2 Data Mining and Knowledge Discovery

This section is a summary of the "data mining and knowledge discovery" virtual panel held at ICDE 2022 on May 10th, 2022. The panelists were listed in alphabetical order as follows:

- Wei Ding (University of Massachusetts, Boston)
- Xuemin Lin (The University of New South Wales, Sydney)
- Jian Pei (Simon Fraser University, Vancouver)

- Shashi Shekhar (University of Minnesota, Minneapolis-St. Paul)
- Xifeng Yan (University of California, Santa Barbara)

The goal of this panel was to gather the world-renowned data mining experts to discuss the achievements and future directions of data mining, which centered around the following questions:

2.1 What are the achievements of data mining in the past 10 years?

The panelists initiated the panel discussion by showing some statistics on publications between 2012 and 2021 that they obtained by searching three keyword phrases "machine learning," "data mining" and "deep learning," for which we use shorthand notations ML, DM and DL hereafter. The findings were that in terms of the number of publications, (1) ML and DL started lower than DM, but now DM lags behind; (2) the United States led till 2019, broke a tie with China in 2020, and China led in 2021; (3) the top-5 most-published institutions in 2021 are University of Chinese Academy of Sciences, Harvard University, Tsinghua University, Shanghai Jiao Tong University, and Stanford University. It is clear that China has made great progress in the last 10 years, but in terms of the number of citations, Stanford University and Harvard University still lead the other institutions by a large margin, indicating higher research impacts.

The panelists also commented that DM has made significant accomplishments in that (1) DM itself has grown into a well-established area in the past 20 years, rather than a sub-field of another area such as database or machine learning; and that (2) DM has become a tool that people would think of for solution when facing real data and real applications. The panelists considered graph mining as one of the fastest-developing fields in DM, witnessing the invention of newer and newer techniques in the past decade such as graph pattern mining, graph embedding, graph neural networks, and the development of knowledge graph as an important sub-field, enabling many real applications such as chemical compound design.

Furthermore, the panelists highlighted spatial data mining as an important achievement in data mining. Spatial data mining is important due to the rise in spatial big data (e.g., remote sensing, census, maps) and important applications such as smart city, climate change, environment, and social good. However, traditional data mining methods face severe challenges in this context. For example, DBSCAN produces spurious patterns in the presence of noise, spatial association rules are unstable due to the modifiable areal unit problem and prediction methods exhibit lower accuracy and spatial bias. To overcome the limitations of one-size-fit-all methods, spatial data mining has provided newer methods (e.g., colocations, significant DBSCAN, spatial autoregression, spatial decision trees, spatial-variability aware neural networks) using the notions of neighbor graphs, spatial autocorrelation, statistical significance testing, etc.

Lastly, the panelists also ranked text mining as an important field in DM that has demonstrated significant progress.

2.2 What are the challenges of data mining?

The panelists pointed out a few challenges for data mining methods. A recent AAAS Science magazine article titled "Taught to Test" [1] said that the current AI (e.g., DL, ML, DM) models perform well on benchmarks but do not generalize well to other (out of sample) datasets. They may embarrass us in real applications due to overfitting to biased benchmarks. For example, the ImageNet benchmark is receiving a lot of negative press because of its racial and gender bias. DL has shown good results in engineering such as solving differential equations, but has much room to improve in other applications such as self-driving cars, where their perception and sensor suites are overwhelmed by adverse weather (e.g., rain, snow, dust). This holds similarly to other conventional DM algorithms. For example, many papers have been published on spatial association rules, but their results are unstable [2], varying dramatically across different choices for space partitioning for a given geospatial dataset. As another example, the DBSCAN clustering algorithm may generate spurious results when there is background

noise, the consequence of which can be very costly in applications such as finding crime or disease hotspots. Furthermore, traditional prediction methods often assume that the data samples are drawn independently of each other and from identical distributions. These assumptions are violated by spatial data, which often exhibit spatial autocorrelation and spatial variability.

Most of the panelists agreed that the number of publications, citations and h-indexes are imperfect measures of impact. For example, based on these metrics, Alan Turing (and about half of the Turing Award winners [3]) would not make it in the list of top 1000 computer scientists. These metrics are not fair to people working in industry and government laboratories, which may prefer patents, products and policy innovations over conference publications. Even for academics, these metrics are not normalized for community size and are not fair to pioneering work in less-crowded but highly important fields. Furthermore, these metrics do not distinguish between self-citation, community citations, conference paper citations, and journal paper citations. Thus, they emphasized the importance of evaluating a DM research based on its transformative impact whether foundational or societal. For example, social-media (e.g., Facebook) recommender algorithms are effective in catching people's attention, but they help spread misinformation and polarize the society, which is undesirable. There is a new trend of responsible computing, and DM researchers may contribute by doing DM research for social good, on topics such as climate change and fairness in AI.

2.3 What is the role of deep learning in data mining research?

Deep Learning (DL) is one of the fastest-developing areas in the past decade and half, so it would be interesting to consider its impacts versus conventional data mining techniques. The panelists pointed out that an obstacle of applying DL models in real applications is that they give blackbox predictions, while people need explainable rules for decision making. In fact, a recent work from Dr. Xifeng Yan's group, BERTRL (BERT-based Relational Learning), showed that association rules can outperform graph neural networks, and it is a promising solution to combine DL with association rule mining where the former filters out unrelated patterns so that the latter can mine the most important rules. In text mining, concepts such as language models and n-grams can all be regarded as a kind of association rules, except that conventional DM methods require explicit definition of association, while DL methods implicitly encode rules in the networks using embeddings. In principle, we can apply any methods as long as we are able to solve problems for users by delivering explainable rules.

The panelists further commented that DM is different from ML in that DM pays more attention to knowledge discovery from data, and to the communication between data owners, data users and domain experts.

Lastly, the panelists suggested that DL and DM are complementary. For example, DL may pre-process image or video data to identify objects, their types (e.g., vehicles) and trajectories, which are then mined by (spatial) DM techniques to identify patterns such as hotspots and their proximal and distant correlates (e.g., colocations, teleconnections), etc. In other words, DL may extend the reach of DM to image and video data. In addition, DL may be used to construct new features to improve performance of DM classifiers.

2.4 What are the promising future directions of data mining research?

Dr. Wei Ding is currently serving as a program director of National Science Foundation (NSF), and based on her experience, several directions are seeing growing interest and encouraged by NSF, including (1) DM research for Medical and Health Sciences, (2) Fairness in AI such as NSF's FAI program, (3) AI-enabled scientific discovery, and (4) software and hardware codesign for AI research targeting extreme scalability. NSF is interested in a broader range of novel machine learning topics and areas, not just DL; in addition, collaborations with domain scientists and industry partners are important, rather than research works that are only verified on toy datasets.

The panelists also considered text mining as a promising direction, since text data is everywhere and thus easy to obtain for knowledge discovery, and there have been great breakthroughs in NLP such as transformer models like the powerful GPT-3. The panelists envisioned that a knowledge-grounded language model would be

an important direction in text mining research, which encourages the convergence of language models, knowledge representation, and rule-based reasoning.

The panelists highlighted the promise of spatiotemporal data mining due to the rise in valuable spatiotemporal big data (e.g., smartphone trajectories, vehicle on-board-diagnostic data, daily scan of Earth from nano-satellites), foundational DM challenges (e.g., temporal non-stationarity, missing data) and socially important applications such as climate change (e.g., forecast sea-level rise and impact), public safety (e.g., monitoring forest fires, flood prediction), public health (e.g., identifying emerging hotspots and predict spread of infectious diseases), and understanding spatiotemporal patterns of lives.

3 Data Mining and Knowledge Discovery 3

The "Data Mining and Knowledge Discovery 3" track includes 12 interesting papers with topics ranging from recommender systems, intent mining, trajectory and time series analytics, rule learning, forecasting, planning, and IoT. A panel discussion follows the paper presentation, where the panelists consist of experts from both academia and industry, including Peilin Zhao (Tencent AI Lab), Ce Zhang (ETH Zurich), Xue Wang (Alibaba DAMO Academy), Jieming Shi (The Hong Kong Polytechnic University), and Boyang Han (JD Intelligent Cities Research). The panel experts share their visions on future research directions in data mining and knowledge discovery and also outline potential application areas that may significantly benefit from such research, summarized as follows.

First, many industrial sectors are undergoing digitalization transformation, providing plenty of unique opportunities for data mining and knowledge discovery (DMKD) research. Novel DMKD techniques will play a more significant role during the transformation. For example, DMKD techniques are now playing a significant role in drug discovery, especially in the stages before pre-clinical trials. In addition, being able to extract data from knowledge is essential for enhancing current intelligent city management, where knowledge fusion, reusing, and transferring are key techniques.

Second, compared to inventing new and better DMKD techniques, how to ensure DMKD techniques accessible and affordable to end users is also important. For instance, time series exist in a wide variety of application domains, and different types of time series analytics models also exist. It can be challenging for domain users to make the right decisions on the analytics models. Thus, automated processes for deploying the most appropriate DMKD techniques for different application domains are called for. In addition, means to reduce computational cost and operational cost are also important.

We believe that data mining and knowledge discovery will continue to be a very hot research area in academia and we will also face many challenging issues when deploying the research results in real world industrial settings, which will inspire further innovation. Thus, we expect to see more and closer collaborations between academia and industry.

4 Federated Learning

In ICDE 2022, the "Federated Learning" track includes 16 interesting research papers. Recently, federated learning (FL) has raised significant attention in both academic and industrial communities [4–6]. In contrast to traditional training paradigm, federated learning is a distributed model training paradigm that enables learning private data knowledge by communicating local model updates rather than gathering the raw data. Serving as an efficient learning scheme for communication and privacy protection, FL has shown its potential to facilitate real-world applications, such as pre-training [7], object detection [8], bio-metrics [9], medical image analysis [10], healthcare [11, 12], finance [13], smart manufacturing [4], and others.

Although FL has demonstrated empirical success in handling the system heterogeneity and data heterogeneity challenges, and there are still many key open issues need to be explored.

- *Federated online learning:* As artificial intelligence IoT (AIoT) devices have been widely deployed in our daily life, a large-scale streaming data needs to be analysis. Most of the existing FL algorithms performs well on statistic analysis. How do design an efficient FL to support online learning is becomes a crucial challenge in AIoT domain.
- *Incentive mechanism design:* The performance of FL usually benefits from more participants joining the training. How to encourage more devices to provide the trained local model is a very interesting direction.
- *Fairness in Federated Learning:* The performance of the global model on each local device are usually different. How to train a fair classifier on decentralized data is very important for each local client.
- *Block-chain in Federated Learning:* As FL suffers from shortcomings such as single-point-failure and malicious data, block-chain provides a secure and efficient solution for the deployment of FL. Currently, a block-chain based federated Learning system in used in Web 3.0 scenarios.

Besides the federated learning, federated database systems are also caught lots of attention by the privacy-preserving features [14, 15]. Since the constituent database systems remain autonomous, a federated database system is a contrastable alternative to the task of merging several disparate databases [16]. The federated database system is basically used when there is some global view or schema of the federation of the database which is basically shared by the applications. The heterogeneity issues of federated database system are differences in data model and data conflicts.

5 Graph Data Management

In ICDE 2022 Graph Data Management session, eleven papers about graph management and analysis were presented, covering a wide range of topics, including: distributed graph analytics, knowledge graphs, graph pattern and isomorphism, influence maximization, and graph learning. Moreover, we have invited four experts in this area:

- Prof. Byron Choi (Hong Kong Baptist University);
- Prof. Arijit Khan (Aalborg University, Denmark);
- Prof. Sibo Wang (Chinese University of Hong Kong); and
- Prof. Yixiang Fang (Chinese University of Hong Kong Shenzhen)

They share their invaluable views on the following two questions below:

5.1 What makes you feel interested/excited in working in this area?

The experts think that graph problems are interesting because they enable the integration of theories and practice. Byron said that there are lot of computational challenges, in particular the need to provide easy-to-use querying tools. Arijit said that graphs are inherent in many areas and systems in the industry, and each of them have domain-specific problems. Sibo explained that it is interesting to develop practial linear-time or sublinear solutions for large graphs, as well as studying multi-core and distributed computing paradigm. Yixiang likes graph problems because the solutions can be used in real graph data.

5.2 What is the most important problem in big graph data in the next 5-10 years?

Byron pointed out that given the variety of graph databases, there is lack of standard for graph query interfaces, so it is crucial to provide standardization for query languages and interfaces. It is also important to develop intuitive querying tools to explore graph data. Arijit followed up by pointing out that graph systems should be designed to allow users to understand graph query answers, and provide feedback to the system, to allow interaction between users and systems. In Sibo's view, the next problem is machine learning on graphs, which involves defining abstractions and atomic operations for graph learning. This can only be done by leveraging expertise in graph algorithms, graph learning, and graph systems. Yixiang pointed out that the next problems involves the study of the building and use of complex networks (e.g., knowledge graphs and multi-graphs) in downstream applications.

There are also active discussions among the audience, raising questions about the user-friendliness problem of existing graph query languages such as SPARQL. Experts responded by explaining that NLP and keyword search may provide better graph management and visualization tools. It is also important to be able to provide explanation to the query answers returned. Among the audience, Prof. Ashraf Aboulnaga said that their group has developed a programmer-friendly option to access RDF graphs, called *RDFFrames* [17]. To conclude, the session gives a nice overview of the latest development of graph management. We would also like to thank the experts for their effort for suggesting their visions about the important and emerging research problems in the next few years.

6 Graph Neural Networks

In ICDE 2022 Graph Nerual Networks session, a number of papers on graph neural networks (GNNs) were presented, ranging from model design, analysis and applications of GNNs. Moreover, five panel experts in this area extensively discussed the research on GNNs.

- Dr. Yixiang Fang (Session chair, The Chinese University of Hong Kong, Shenzhen);
- Dr. Xiao Wang (Beijing University of Posts and Telecommunications);
- Dr. Rong Zhu (Alibaba DAMO Academy);
- Dr. Jieming Shi (The Hong Kong Polytechnic University);
- Dr. Wenqing Lin (Tencent).

In particular, they have extensively discussed the following five future research directions on GNN:

6.1 Theoretical foundation of GNNs

Unlike traditional graph mining models, GNNs provide an automatic paradigm to summarize structure and label information of graph data into hidden representations. However, the boundary of the expressive power of GNNs is still unknown. Existing works try to analyze GNNs' expressive power using Weisfeiler-Lehman isomorphism test. The 1-hop neighbor based GNNs (e.g., GraphSage, GAT and GCN) cannot do many simple tasks (e.g., triangle counting). It is still unclear whether complex forms of GNNs could mitigate this gap. Only when we develop more powerful tools to analyze the capacity boundary of GNNs, we can better know the strengths and weaknesses of GNNs compared to previous models.

6.2 Interpretability of GNNs

How to explain the learned embedding vectors of GNNs is another challenging task. Unlike traditional models, GNNs cannot tell which motif and/or which label plays important roles in the resulting knowledge. This prevents GNNs to be used in some scenarios that are with serious effect, such as medicine and financial applications. Therefore, it is crucial to understand the underlying mechanisms of the models in human terms.

6.3 Killer applications of GNNs

Although GNN has received much attention recently, we have not witnessed any killer application that must rely on GNNs. Unlike PageRank which lays the foundation of web page search engine, GNN has not proved its importance in such a widely application scenario. We admit GNNs have advantages in automatic learning, but still lack killer applications. Thus, we should calm down to think deeper on the importance and roles of GNNs in more real applications.

6.4 Efficiency issues of GNNs

Nowadays, big graphs are prevalent in various areas. Nevertheless, most of existing GNN models cannot process large graphs at scale. Consequently, how to build GNN models that achieve both high efficiency and strong scalability without sacrificing accuracy on large graphs is an important future research direction.

6.5 Trustworthy GNNs

Current GNNs still mainly focus on the performance improvement. However, when we deploy GNNs to the real world scenarios, especially some risk-sensitive areas, the accuracy is not the only metric to evaluate the GNNs. Whether the GNNs are trustworthy is an important factor. For example, how to generalize the GNNs to the out-of-distribution graphs, i.e., how can we ensure that the performance of GNNs keeps stable when the distribution of test graphs changes. Besides, the robustness and fairness of GNNs are also very important in real applications.

In addition, many other related issues of GNN, including general frameworks, auto-search of parameters, benchmarks, and platforms/systems, have also been briefly discussed.

7 Spatial and Temporal Data Management

The 38th IEEE International Conference on Data Engineering (ICDE 2022) was held in May 9-12, 2022 as a virtual event. ICDE2022 Session “Spatial and Temporal Data Management” was chaired by Nikos Mamoulis and ran between 14:30 and 16:10 Malay time. The panelists include six experts in this area:

- Dr. Panagiotis Bouros (Johannes Gutenberg University)
- Dr. Anton Dignös (Free University of Bozen-Bolzano)
- Dr. Nikos Mamoulis (University of Ioannina)
- Dr. Matthias Renz (CAU University of Kiel)
- Dr. Raymond Chi-Wing Wong (The Hong Kong University of Science and Technology)
- Dr. Eleni Tzirita Zacharatou (IT University of Copenhagen)

The session included 11 papers, naturally partitioned into 4 sub-sessions, as follows:

Spatial Data:

- A Machine Learning-Aware Data Re-partitioning Framework for Spatial Datasets
- Example-based Spatial Search at Scale
- SPADE: GPU-Powered Spatial Database Engine for Commodity Hardware

Spatial Crowdsourcing:

- Bilateral Preference-aware Task Assignment in Spatial Crowdsourcing
- Human-Drone Collaborative Spatial Crowdsourcing by Memory-Augmented Distributed Multi-Agent Deep
- Influence-aware Task Assignment in Spatial Crowdsourcing

Trajectory Data:

- Maximizing Range Sum in Trajectory Data
- Workload-Aware Shortest Path Distance Querying in Road Networks

Temporal and Time-Series Data:

- Provenance in Temporal Interaction Networks
- Constructing Compact Time Series Index for Efficient Window Query Processing
- iTemporal: An Extensible Generator of Temporal Benchmarks

Papers in the first sub-session apply data preprocessing to assist spatial ML tasks, study spatial pattern queries and use modern hardware to accelerate spatial data management. Papers about spatial crowdsourcing studied how to incorporate more information (e.g., drone data, influence of objects and preferences from people) for matching tasks with human in a more realistic setting. Papers in the third sub-session, study problems for road-network applications (i.e., range sum maximization over trajectory data, shortest path queries). Papers in the last sub-session are on various topics related to temporal and time-series management (i.e., computing provenance information in temporal networks, time-series indexing, temporal benchmark generation).

The authors of each paper took about 5 minutes to present their work. After the end of all presentations, the session chair coordinated a discussion between the audience, the invited panel members, and the authors, which included Q&A and a discussion about future directions on the topics.

7.1 Categorization and Summary of the Papers

Based on their focus, we could categorize the papers of the session into the following classes:

1. Application-oriented paper(s)
2. Paper(s) on Database Indexing
3. Paper(s) on Database Benchmarking
4. Paper(s) on Special Topics that involve Spatial, Temporal and Spatiotemporal Data

This categorization gives us insights for trends and possible future directions for spatial and/or temporal data engineering. Hence, one direction is to focus on real-life applications that manage spatial/temporal data, in order to devise effective and efficient solutions for their needs. The second direction is indexing on spatial/temporal data, which is a fundamental problem in our community. The third direction is to provide effective and reliable benchmarks for spatial/temporal data management. The fourth direction is working on research about combination of other research topics with spatial/temporal databases. Such topics include machine learning, human involvement, graph search, and new technologies.

In addition to these four directions that stem directly from the categorization, panel members discussed further research directions which are outlined in the following sections.

7.2 Spatial and Temporal Data Fusion

The increased availability of data and knowledge as well as the trend towards more interdisciplinary and transdisciplinary research approaches call for solutions unlocking the immense power of cross domain and multi-source data fusion, i.e. the fusion of views representing the various perspectives on a scene of an excerpt of the real world. Thereby, views vary among different disciplines, different data sets, data types and data sources including models, and different abstraction of data ranging from measurement records over patterns up to knowledge representations. For example, physicists and biologists have different perspectives on some scenes in the ocean, such as the behavior of saline concentration and behaviour of living organisms. Also within the same scientific discipline, there may be many varying views on entities of a general research topic. For example oxygen deficits in coastal regions measured by remote sensing raster data vs. series of in-situ sensor measurements. Since most scientific views are traditionally explored in isolation, they are also restricted to specific findings and often do not allow a broad understanding of relationships and functional dependencies between multiple concepts such as structural relationships among different scientific views. It was coined that research on heterogeneous data integration and fusion for the spatial and temporal data has not been studied extensively in the literature. This creates many opportunities for research. To pave the path towards cross-domain Fusion, a novel, fundamental and systematic framework of methods is needed that enable the fusion of data, patterns, knowledge, etc.; i.e., the fusion of multiple potentially heterogeneous views and stages of data from diverse domains. Initial approaches for exploiting the power of machine learning to fusing data/views lifting up the potential of data analysis are summarized in [18]. While most approaches have been proposed in the field of urban analytics, fusing traffic data with other urban attributes [19–21], it increasingly attracts further scientific disciplines [22, 23]. However, most existing solutions are isolated, often too narrow, ad-hoc approaches designed for specific applications. We need more general, broader and systematic fusion concepts.

With the integration of diverse data sources, we also need solutions for the alignment of diverse types and models of data, in particular spatial and temporal data occurs in different forms, scales and structures. In heterogeneous data sources but also due to its particular nature, temporal data occurs in different forms. Interval data is used to store state information over a period of time, e.g., an employment contract or a task allocated on a manufacturing machine, event data is used to store a happening at a specific point in time, e.g., a warning issued to an employee or a failure of a machine component, and time series data is used to store continuous variables or parameters, e.g., stress levels of employees or sensor data related to manufacturing machines. While processing and analysis techniques exist for temporal data, they usually focus on a single type, and do not consider the fusion of the different form based on their nature, resulting in an incomplete picture for analysis. This, despite the fact that different types of temporal data arise in the same application, for instance in [24] they consider benchmarks that generate both time intervals and time points. In this setting, there is a need for research on data modeling for being able to reference data entries based on their type and timestamp that belong together, research on data processing for the efficient fusion of data entries that need to be analyzed together, and finally, analysis techniques that consider all three types of temporal data.

Examples of interval, event and time series data from our “Spatial and Temporal Data Management” session.

In [25] the input data is time series data, while the window sequence is interval data; in [24] they consider the generation of interval data and event data or time series data; in [26] they consider event data and the windowing approach (maybe also used for tracking over particular interesting time period) employs interval data.

7.3 Plug-and-Play Infrastructure

Most of the existing work in spatial data management, including papers in this "Spatial and Temporal Data Management" session such as [27, 28], focuses on devising high-performance solutions to specific problems. However, looking at the big picture, this is insufficient. In addition to those specialized tailor-made solutions, we need to create an infrastructure that supports a "plug-and-play" functionality. Specifically, this infrastructure should enable and ease the creation and composition of processing pipelines that leverage individual specialized advances. In addition, the envisioned infrastructure should serve as a bridge between data management researchers and domain experts. While today we have abundant access to spatio-temporal data, it is often hard to determine what is interesting and challenging about this data without the knowledge of a domain expert. Therefore, to foster interdisciplinary collaborations, the envisioned infrastructure should facilitate a feedback loop between data management researchers and domain experts.

7.4 Spatial Data Management Beyond Vector Data

All the papers in the spatial data management sub-session [27–29] deal with objects that follow the vector data model, which represents spatial features with their geometry. Typical geometries are points (e.g., GPS locations), lines (e.g., roads and rivers) and polygons (e.g., regional boundaries). Nowadays, Earth Observation (EO) satellites are another prominent source of spatial data. The number of EO satellites and their acquisition rates are growing fast, resulting in an ever-increasing amount of collected satellite images. Satellite images follow the raster data model, which represents the study area as a collection of granules (e.g., pixels). To enable efficient data analyses, we need data management solutions that treat both data types as first-class citizens. For example, tracking flooded residential areas during an ongoing flood event requires comparing water masks from recent satellite images (rasters) with permanent water bodies (vectors). While there is some initial work towards the efficient combination of vector and raster data [30], many open challenges remain in query execution and optimization.

7.4.1 Spatial Query Optimization and Performance Tuning

Today, we observe a twofold evolution of spatial data systems and algorithms. From the one side, the application domains become more and more diverse [27, 28, 31, 32], while systems incorporate new hardware technologies in their design [29]. This twofold evolution creates new challenges in query optimization and performance tuning. For example, the query optimizer proposed in [29] only considers the cost of the data transfer to the GPU and lacks support for multiway queries and queries with multiple constraints. Further research is required to develop query optimization and performance tuning techniques that can navigate a broad spectrum of application requirements and hardware settings.

7.5 Privacy-Aware Compliant Spatial Crowdsourcing

The "Spatial and Temporal Data Management" session contained several papers that advance the state-of-the-art in spatial crowdsourcing [33–35]. However, enriching existing solutions with privacy awareness and compliance with regulatory restrictions (for example, with respect to the use of drones proposed in [34]) remains an open research challenge.

7.6 Road Networks and Trajectory Data

Road networks are spatial graphs used in a variety of real-life applications. Routing being one of them, lies in the core of navigation systems, location-aware recommender systems for touristic and trip-planning scenarios and logistics, among others. In its simplest and most fundamental form, the goal is to determine the shortest or the fastest path between two locations in the network. A number of path-finding algorithms and indexing techniques has been proposed in the past to efficiently process shortest paths queries. Performance and scalability are traditionally the interesting objectives, while dealing with structural updates is less important as road networks are very static, contrary to other types of graphs. In the dynamic aspect of routing, computing time-aware paths is a timely challenge, where how to consider moving patterns in different times of the day and employ real-time traffic are key challenges. With the proliferation of geo-positioning systems and mobile phones, the amount of routing queries received by mobile applications has significantly increased. Under this, techniques for parallel and batch processing such queries are important, but another interesting research question is how we can use previous requests (i.e., a workload of routing queries) to optimize existing path-finding techniques and indices. [32] focuses on how the spatial skewness of the query locations and the temporal locality of the time-aware requests in an existing workload can be considered towards this direction.

Besides the actual graph of a road network, trajectories are another type of data used by modern location-aware applications. In routing, trajectories can be used as train data for applying learning techniques to improve the quality of the recommended paths. For example, previously followed paths by users or their friends can be used to learn moving habits and patterns. Trajectories play a key role also in business placement scenarios where the goal is to identify the best location for a company to open a store. Such scenario is studied in [31].

8 Spatial and Temporal Data Mining

In ICDE 2022, the “Spatial and Temporal Data Mining" session contains 12 research papers. The panelists include five experts in this area:

- Dr. Zhifeng Bao (RMIT University)
- Dr. Jilin Hu (Aalborg University)
- Dr. Cheng Long (Nanyang Technological University)
- Dr. Raymond Chi-Wing Wong (The Hong Kong University of Science and Technology)
- Dr. Bin Yang (Aalborg University)

Spatial and temporal data, which refers to data that involves spatial information (e.g., coordinates) and/or temporal information (e.g., time stamps), is continuously being generating in a wide range of applications in urban cities (e.g., mobilities, commuting, traffic, planning and logistics), in geography (e.g., GIS and remote sensing), in chemistry and biology (e.g., 3D molecular modeling), etc. Mining spatial and temporal data in these applications would naturally facilitate these applications with more intelligence, effectiveness and/or efficiency. At ICDE 2022, we have a dozen interesting papers published in this area of spatial and temporal data mining. We also have a few experts, sharing their thoughts and visions of conducting research in this area. Next, we first outline these ICDE 2022 papers in this area and then summarize the expert panels' thoughts and visions of this area.

The papers fall in three groups, namely (1) time series mining [36–40], (2) spatiotemporal prediction [41–45] and (3) representation learning for spatial and/or temporal data [46, 47]. Specifically, [36, 37] propose autoencoder networks for time series outlier detection, [38, 39] study shapelets of time series for tasks such as time series classification, and [40] studies the backdoor attack on deep learning based time series classification

models, [41–45] study various spatiotemporal prediction tasks (including activity prediction [41], urban crime prediction [42], traffic prediction [43], docked bike prediction [44]) and the problem of setting the grid size for spatiotemporal prediction models [45], and [46, 47] study the presentation learning problem for temporal paths [46] and trajectory [47].

The panel experts share their visions and suggestions of conducting research in spatial and temporal data mining for the next 5-10 years, which are summarized as follows. First, this area of research will continue to be attractive with the new developments of technologies such as AI, 5G and metaverse. For example, metaverse corresponds to a virtual physical space which naturally involves spatial data such as terrain data. Second, researchers in this area can be more aggressive by taking a leading role in developing new technologies instead of adopting those from other areas such as CNNs, GNNs, and word2vec. Third, developing generic frameworks that are able to enhance a variety of existing models rather than yet another end-to-end model would make more impact. Some examples of such generic frameworks include for turning ST-agnostic models to ST-aware models and [48, 49] frameworks for automatically searching for models of time series forecasting. Fourth, collaborating with industry more closely and conducting interdisciplinary research involving spatial and temporal data would have significant potential and bring real impact. Fifth, existing spatial and temporal data mining solutions still need to be improved in various aspects, including but limited to interpretability, scalability, sustainability, data sparsity and uncertainty. All in all, we believe that spatial and temporal data mining will continue to be a hot research area and involve many research problems to be solved ahead of the way.

9 Conclusion

In conclusion, the introduction of panel discussions at the 38th IEEE International Conference on Data Engineering (ICDE) proved highly successful in generating inspiring discussions and promoting active participation. These discussions enriched understanding, offered fresh insights, and attracted more participants than anticipated. Summarizing the panels in this article benefits the data engineering community, providing a valuable resource for researchers and conference organizers. The success of the panel discussions at ICDE'22 highlights their effectiveness in fostering engagement and collaboration. The article suggests adopting this format for future conferences, encouraging in-depth exploration of research topics and active involvement from diverse perspectives.

Acknowledgement: This article was compiled by the PC co-chairs (Gao Cong, Stratos Idreos, and Feifei Li) of ICDE'22.

References

[1] Matthew Hutson. Taught to the Test. IEEE Aerospace and Electronic Systems Magazine, Vol 36, 9, 30-42, 2021.

[2] Eftelioglu, Emre and Shekhar, Shashi and Hudson, James and Joppa, Lucas and Baru, Chaitanya and Janeja, Vandana. Data Science for Earth: An Earth Day Report. Association for Computing Machinery, Vol 22, 1, 4-7, 2020.

[3] Jin, Yinyu and Yuan, Sha and Zhou, Shao and Hall, Wendy and Tang, Jie. Turing Award elites revisited: patterns of productivity, collaboration, authorship and impact. Scientometrics, Vol 126, 2021.

[4] Hao, Meng and Li, Hongwei and Luo, Xizhao and Xu, Guowen and Yang, Haomiao and Liu, Sen. Efficient and privacy-enhanced federated learning for industrial artificial intelligence. IEEE Transactions on Industrial Informatics, arXiv preprint, 2019.

[5] Li, Tian and Sahu, Anit Kumar and Talwalkar, Ameet and Smith, Virginia. Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, Vol 37, 3, 2020.

[6] Nguyen, Dinh C and Ding, Ming and Pathirana, Pubudu N and Seneviratne, Aruna and Li, Jun and Poor, H Vincent. Federated learning for internet of things: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2021.

[7] Tian, Yuanyishu and Wan, Yao and Lyu, Lingjuan and Yao, Dezhong and Jin, Hai and Sun, Lichao. FedBERT: When Federated Learning Meets Pre-Training. ACM Transactions on Intelligent Systems and Technology (TIST), 2022.

[8] Liu, Yang and Huang, Anbu and Luo, Yun and Huang, He and Liu, Youzhi and Chen, Yuanyuan and Feng, Lican and Chen, Tianjian and Yu, Han and Yang, Qiang. Fedvision: An online visual object detection platform powered by federated learning. AAAI, 2020.

[9] Dayan, Ittai and Roth, Holger R and Zhong, Aoxiao and Harouni, Ahmed and Gentili, Amilcare and Abidin, Anas Z and Liu, Andrew and Costa, Anthony Beardsworth and Wood, Bradford J and Tsai, Chien-Sung and others. Federated learning for predicting clinical outcomes in patients with COVID-19. Nature medicine, Vol 27, 10, 1735-1743, 2021.

[10] Kaissis, Georgios A and Makowski, Marcus R and Rückert, Daniel and Braren, Rickmer F. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, Vol 2, 6, 305-311, 2020.

[11] Xu, Jie and Glicksberg, Benjamin S and Su, Chang and Walker, Peter and Bian, Jiang and Wang, Fei. Federated learning for healthcare informatics. Journal of Healthcare Informatics Research, Vol 5, 1, 2021.

[12] Antunes, Rodolfo Stoffel and André da Costa, Cristiano and Küderle, Arne and Yari, Imrana Abdullahi and Eskofier, Björn. Federated Learning for Healthcare: Systematic Review and Architecture Proposal. ACM Transactions on Intelligent Systems and Technology (TIST), Vol 22, 4, 1-23, 2022.

[13] Yang, Wensi and Zhang, Yuhang and Ye, Kejiang and Li, Li and Xu, Cheng-Zhong. Ffd: A federated learning based method for credit card fraud detection. International conference on big data, 2019.

[14] Larson, James A and Larson, Carol L. Federated Database Systems. Handbook of Heterogeneous Networking 1999, 27-1, 2018.

[15] Gupta, Ankush M and Gadepally, Vijay and Stonebraker, Michael. Cross-engine query execution in federated database systems. IEEE High Performance Extreme Computing Conference (HPEC), 1-6, 2016.

[16] Sheth, Amit P and Larson, James A. Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys (CSUR), Vol 22, 3, 183-236, 1990.

[17] Aisha Mohamed, Ghadeer Abuoda, Abdurrahman Ghanem, Zoi Kaoudi, Ashraf Aboulnaga. RDFFrames: knowledge graph access for machine learning tools. VLDB J., Vol 31, 2, 321-346, 2022.

[18] Meng, Tong and Jing, Xuyang and Yan, Zheng and Pedrycz, Witold. A survey on machine learning for data fusion. Information Fusion, Vol 57, 115-129, 2020.

[19] Liu, Jia and Li, Tianrui and Xie, Peng and Du, Shengdong and Teng, Fei and Yang, Xin. Urban big data fusion based on deep learning: An overview. Information Fusion, Vol 53, 123-133, 2020.

[20] Khan, Sulaiman and Nazir, Shah and García-Magariño, Iván and Hussain, Anwar. Deep learning-based urban big data fusion in smart cities: Towards traffic monitoring and flow-preserving fusion. Computers & Electrical Engineering, Vol 89, 106906, 2021.

[21] Zheng, Yu. Methodologies for cross-domain data fusion: An overview. IEEE transactions on big data, Vol 1, 1, 16-34, 2015.

[22] Soldi, Giovanni and Gaglione, Domenico and Forti, Nicola and Millefiori, Leonardo M and Braca, Paolo and Carniel, Sandro and Di Simone, Alessio and Iodice, Antonio and Riccio, Daniele and Daffinà, Filippo Cristian and others. Space-based global maritime surveillance. Part II: Artificial intelligence and data fusion techniques. IEEE Aerospace and Electronic Systems Magazine, Vol 36, 9, 30-42, 2021.

[23] Karagiannopoulou, Aikaterini and Tsertou, Athanasia and Tsimiklis, Georgios and Amditis, Angelos. Data Fusion in Earth Observation and the Role of Citizen as a Sensor: A Scoping Review of Applications, Methods and Future Trends. Remote Sensing, Vol 14, 5, 1263, 2022.

[24] Luigi Bellomarini, Markus Nissl, Emanuel Sallinger. iTemporal: An Extensible Generator of Temporal Benchmarks. IEEE ICDE, 2022-2034, 2022.

[25] Jing Zhao. Peng Wang, Bo Tang, Lu Liu, Chen Wang, Wei Wang, Jianmin Wang. Constructing Compact Time Series Index for Efficient Window Query Processing. IEEE ICDE, 3025-3037, 2022.

[26] Chrysanthi Kosyfaki, Nikos Mamoulis. Provenance in Temporal Interaction Networks. IEEE ICDE, 2278-2291, 2022.

[27] Kanchan Chowdhury, Venkata Vamsikrishna Meduri, Mohamed Sarwat. A Machine Learning-Aware Data Re-Partitioning Framework for Spatial Datasets. IEEE ICDE, 2427-2440, 2022.

[28] Hanyuan Zhang, Siqiang Luo, Jieming Shi, Jing Nathan Yan, Weiwei Sun. Example-Based Spatial Search at Scale. IEEE ICDE, 539-551, 2022.

[29] Harish Doraiswamy, Juliana Freire. SPADE: GPU-Powered Spatial Database Engine for Commodity Hardware. IEEE ICDE, 2670-2682, 2022.

[30] Samriddhi Singla, Ahmed Eldawy, Tina Diao, Ayan Mukhopadhyay, Elia Scudiero. The Raptor Join Operator for Processing Big Raster + Vector Data. SIGSPATIAL, 324-335, 2021.

[31] Kaiqi Zhang, Hong Gao, Xixian Han, Jian Chen, Jianzhong Li. Maximizing Range Sum in Trajectory Data. IEEE ICDE, 755-766, 2022.

[32] Bolong Zheng, Jingyi Wan, Yongyong Gao, Yong Ma, Kai Huang , Xiaofang Zhou, Christian S. Jensen. Workload-Aware Shortest Path Distance Querying in Road Networks. IEEE ICDE, 2373-2385, 2022.

[33] Xuanhao Chen, Yan Zhao, Kai Zheng, Bin Yang, Christian S. Jensen. Influence-Aware Task Assignment in Spatial Crowdsourcing. IEEE ICDE, 2142-2154, 2022.

[34] Yu Wang, Chi Harold Liu, Chengzhe Piao, Ye Yuan, Rui Han, Guoren Wang, Jian Tang. Human-Drone Collaborative Spatial Crowdsourcing by Memory-Augmented Distributed Multi-Agent Deep Reinforcement Learning. IEEE ICDE, 459-471, 2022.

[35] Xu Zhou, Shiting Liang, Kenli Li, Yunjun Gao, Keqin Li. Bilateral Preference-Aware Task Assignment in Spatial Crowdsourcing. IEEE ICDE, 1688-1700, 2022.

[36] Tung Kieu, Bin Yang, Chenjuan Guo, Razvan-Gabriel Cirstea, Yan Zhao, Yale Song, Christian S. Jensen. Anomaly Detection in Time Series with Robust Variational Quasi-Recurrent Autoencoders. IEEE ICDE, 1342-1354, 2022.

[37] Tung Kieu, Bin Yang, Chenjuan Guo, Christian S. Jensen, Yan Zhao, Feiteng Huang, Kai Zheng. Robust and Explainable Autoencoders for Unsupervised Time Series Outlier Detection. IEEE ICDE, 3038-3050, 2022.

[38] Guozhong Li, Byron Choi, Jianliang Xu, Sourav S. Bhowmick, Daphne Ngar-yin Mah, Grace Lai-Hung Wong. IPS: Instance Profile for Shapelet Discovery for Time Series Classification. IEEE ICDE, 1781-1793, 2022.

[39] Akihiro Yamaguchi, Ken Ueno, Hisashi Kashima. Learning Evolvable Time-series Shapelets. IEEE ICDE, 793-805, 2022.

[40] Daizong Ding, Mi Zhang, Yuanmin Huang, Xudong Pan, Fuli Feng, Erling Jiang , Min Yang. Towards Backdoor Attack on Deep Learning based Time Series Classification. IEEE ICDE, 1274-1287, 2022.

[41] Yinfeng Li, Chen Gao, Quanming Yao, Tong Li, Depeng Jin, Yong Li. DisenHCN: Disentangled Hypergraph Convolutional Networks for Activity Prediction. CoRR, abs/2208.06794, 2022.

[42] Zhonghang Li, Chao Huang, Lianghao Xia, Yong Xu, Jian Pei. Spatial-Temporal Hypergraph Self-Supervised Learning for Crime Prediction. IEEE ICDE, 2984-2996, 2022.

[43] Razvan-Gabriel Cirstea, Bin Yang, Chenjuan Guo, Tung Kieu, Shirui Pa. Towards Spatio- Temporal Aware Traffic Time Series Forecasting. IEEE ICDE, 2900-2913, 2022.

[44] Guanyao Li, Xiaofeng Wang, Gunarto Sindoro Njoo, Shuhan Zhong, S.-H. Gary Chan, Chih-Chieh Hung, Wen-Chih Peng. A Data-Driven Spatial-Temporal Graph Neural Network for Docked Bike Prediction. IEEE ICDE, 713-726, 2022.

[45] Jiabao Jin, Peng Cheng, Lei Chen, Xuemin Lin, Wenjie Zhang. GridTuner: Reinvestigate Grid Size Selection for Spatiotemporal Prediction Models. IEEE ICDE, 1193-1205, 2022.

[46] Sean Bin Yang, Chenjuan Guo, Jilin Hu, Bin Yang, Jian Tang, Christian S. Jensen. Weakly-supervised Temporal Path Representation Learning with Contrastive Curriculum Learning. IEEE ICDE, 2873-2885, 2022.

[47] Yang, Peilun and Wang, Hanchen and Lian, Defu and Zhang, Ying and Qin, Lu and Zhang, Wenjie. TMN: Trajectory Matching Networks for Predicting Similarity. IEEE ICDE, 1700-1713, 2022.

[48] Razvan-Gabriel Cirstea, Chenjuan Guo, Bin Yang, Tung Kieu, Xuanyi Dong, Shirui Pan. Triformer: Triangular, Variable-Specific Attentions for Long Sequence Multivariate Time Series Forecasting. IJCAI, 1994-2001, 2022.

[49] Razvan-Gabriel Cirstea, Tung Kieu, Chenjuan Guo, Bin Yang, Sinno Jialin Pan. EnhanceNet: Plugin Neural Networks for Enhancing Correlated Time Series Forecasting. IEEE ICDE, 1739-1750, 2021.