SIGMOD Officers, Committees, and Awardees

Chair
Divyakant Agrawal
Department of Computer Science
UC Santa Barbara
Santa Barbara, California
USA
+1 805 893 4385
agrawal <at> cs.ucsb.edu

Vice-Chair
Fatma Ozcan
Systems Research Group
Google
Sunnyvale, California
USA
+1 669 264 9238
Fozcan <at> google.com

Secretary/Treasurer
Rachel Pottinger
Department of Computer Science
University of British Columbia
Vancouver
Canada
+1 604 822 0436
Rap <at> cs.ubc.ca

SIGMOD Executive Committee:

Divyakant Agrawal (Chair), Fatma Ozcan (Vice-chair), Rachel Pottinger (Treasurer), Juliana Freire (Previous SIGMOD Chair), Chris Jermaine (SIGMOD Conference Coordinator), Rada Chirkova (SIGMOD Record Editor), Alexandra Meliou (2024 SIGMOD PC co-chair), S Sudarshan (2024 SIGMOD PC co-chair), Floris Geerts (Chair of PODS), Genoveva Vargas Solar (SIGMOD Diversity and Inclusion Coordinator), Sourav S Bhowmick (SIGMOD Ethics), Yufei Tao (ACM TODS Editor in Chief)

Advisory Board:

Yannis Ioannidis (Chair), Phil Bernstein, Surajit Chaudhuri, Rakesh Agrawal, Joe Hellerstein, Mike Franklin, Laura Haas, Renee Miller, John Wilkes, Chris Olsten, AnHai Doan, Tamer Özsu, Gerhard Weikum, Stefano Ceri, Beng Chin Ooi, Timos Sellis, Sunita Sarawagi, Stratos Idreos, and Tim Kraska

SIGMOD Information Directors:

Sourav S Bhowmick, Nanyang Technological University
Byron Choi, Hong Kong Baptist University

Associate Information Directors:

Hui Li (SIGMOD Record), Georgia Koutrika (Blogging), Wim Martens (PODS)

SIGMOD Record Editor-in-Chief:

Rada Chirkova, NC State University

SIGMOD Record Associate Editors:

Lyublena Antova, Manos Athanassoulis, Angela Bonifati, Renata Borovica-Gajic, Vanessa Braganholo, Aaron J. Elmore, George Fletcher, Wook-Shin Han, H V Jagadish, Alfons Kemper, Benny Kimelfeld, Samuel Madden, Kyriakos Mouratidis, Tamer Özsu, Kenneth Ross, Pınar Tözün, Immanuel Trummer, Yannis Velegrakis, and Ke Yi

SIGMOD Conference Coordinator:

Chris Jermaine, Rice University

PODS Executive Committee:

Floris Geerts (chair), Pablo Barcelo, Leonid Libkin, Hung Q. Ngo, Reinhard Pichler, Dan Suciu

Sister Society Liaisons:

Raghu Ramakhrishnan (SIGKDD), Yannis Ioannidis (EDBT Endowment), Christian Jensen (IEEE TKDE)

SIGMOD Awards Committee:

Sharad Mehrotra (Chair), H.V. Jagadish, Sourav S Bhowmick, Angela Bonifati, David Maier, Sayan Ranu, Wang-Chiew Tan

Jim Gray Doctoral Dissertation Award Committee:

Evaggelia Pitoura (chair), Angela Bonifati (co-chair), Sourav S Bhowmick, Daniel Kang, Georgia Koutrika, Supun Nakandala, Fatma Ozcan, Julia Stoyanovich, and Xiaofang Zhou

SIGMOD Edgar F. Codd Innovations Award

For innovative and highly significant contributions of enduring value to the development, understanding, or use of database systems and databases. Recipients of the award are the following:

Michael Stonebraker (1992)
David DeWitt (1995)
Serge Abiteboul (1998)
Rudolf Bayer (2001)
Ronald Fagin (2004)
Jennifer Widom (2007)
Umeshwar Dayal (2010)
Stefano Ceri (2013)
Gerhard Weikum (2016)
Anastasia Ailamaki (2019)
Dan Suciu (2022)
Carlo Zaniolo (2025)

Jim Gray (1993)
C. Mohan (1996)
Hector Garcia-Molina (1999)
Patricia Selinger (2002)
Michael Carey (2005)
Moshe Y. Vardi (2008)
Surajit Chaudhuri (2011)
Martin Kersten (2014)
Goetz Graefe (2017)
Beng Chin Ooi (2020)
Joseph M. Hellerstein (2023)

Philip Bernstein (1994)
David Maier (1997)
Rakesh Agrawal (2000)
Don Chamberlin (2003)
Jeffrey D. Ullman (2006)
Masaru Kitsuregawa (2009)
Bruce Lindsay (2012)
Laura Haas (2015)
Raghu Ramakrishnan (2018)
Alon Halevy (2021)
Samuel Madden (2024)

SIGMOD Systems Award

For technical contributions that have had significant impact on the theory or practice of large-scale data management systems.

Michael Stonebraker and Lawrence Rowe (2015); Martin Kersten (2016); Richard Hipp (2017); Jeff Hammerbacher, Ashish Thusoo, Joydeep Sen Sarma; Christopher Olston, Benjamin Reed, and Utkarsh Srivastava (2018); Xiaofeng Bao, Charlie Bell, Murali Brahmadesam, James Corey, Neal Fachan, Raju Gulabani, Anurag Gupta, Kamal Gupta, James Hamilton, Andy Jassy, Tengiz Kharatishvili, Sailesh Krishnamurthy, Yan Leshinsky, Lon Lundgren, Pradeep Madhavarapu, Sandor Maurice, Grant McAlister, Sam McKelvie, Raman Mittal, Debanjan Saha, Swami Sivasubramanian, Stefano Stefani, and Alex Verbitski (2019); Don Anderson, Keith Bostic, Alan Bram, Grg Burd, Michael Cahill, Ron Cohen, Alex Gorrod, George Feinberg, Mark Hayes, Charles Lamb, Linda Lee, Susan LoVerso, John Merrells, Mike Olson, Carol Sandstrom, Steve Sarette, David Schacter, David Segleau, Mario Seltzer, and Mike Ubell (2020); Michael Blanton, Adam Bolton, Bill Boroski, Joel Brownstein, Robert Brunner, Tamas Budavari, Sam Carliles, Jim Gray, Steve Kent, Peter Kunszt, Gerard Lemson, Nolan Li, Dmitry Medvedev, Jeff Munn, Deoyani Nandrekar-Heinis, Maria Nieto-Santisteban, Wil O'Mullane, Victor Paul, Don Slutz, Alex Szalay, Gyula Szokoly, Manu Taghizadeh-Popp, Jordan Raddick, Bonnie Souter, Ani Thakar, Jan Vandenberg, Benjamin Alan Weaver, Anne-Marie Weijmans, Sue Werner, Brian Yanny, Donald York, and the SDSS collaboration (2021); Michael Armbrust, Tathagata Das, Ankur Dave, Wenchen Fan, Michael J. Franklin, Huaxin Gao, Maxim Gekk, Ali Ghodsi, Joseph Gonzalez, Liang-Chi Hsieh, Dongjoon Hyun, Hyukjin Kwon, Xiao Li, Cheng Lian, Yanbo Liang, Xiangrui Meng, Sean Owen, Josh Rosen, Kousuke Saruta, Scott Shenker, Ion Stoica, Takuya Ueshin, Shivaram Venkataraman, Gengliang Wang, Yuming Wang, Patrick Wendell, Reynold Xin, Takeshi Yamamuro, Kent Yao, Matei Zaharia, Ruifeng Zheng, and Shixiong Zhu (2022); Aljoscha Krettek, Andrey Zagrebin, Anton Kalashnikov, Arvid Heise, Asterios Katsifodimos, Jiangji (Becket) Qin, Benchao Li, Bowen Li, Caizhi Weng, ChengXiang Li, Chesnay Schepler, Chiwan Park, Congxian Qiu, Daniel Warneke, Danny Cranmer, David Anderson, David Morávek, Dawid Wysakowicz, Dian Fu, Dong Lin, Eron Wright, Etienne Chauchot, Fabian Hueske, Fabian Paul, Feng Wang, Gabor Somogyi, Gary Yao, Godfrey He, Greg Hogan, Guowei Ma, Gyula Fora, Haohui Mai, Henry Saputra, Hequn Cheng, Igal Shilman, Ingo Bürk, Jamie Grier, Jark Wu, Jincheng Sun, Jing Ge, Jing Zhang, Jingsong Lee, Junhan Yang, Konstantin Knauf, Kostas Kloudas, Kostas Tzoumas, Kete (Kurt) Young, Leonard Xu, Lijie Wang, Lincoln Lee, Lungu Andra, Martijn Visser, Marton Balassi, Matthias J. Sax, Matthias Pohl, Matyas Orhidi, Maximilian Michels, Nico Kruber, Niels Basjes, Paris Carbone, Piotr Nowojski, Qingsheng Ren, Robert Metzger, Roman Khachatryan, Rong Rong, Rui Fan, Rui Li, Sebastian Schelter, Seif Haridi, Sergey Nuyanzin, Seth Wiesman, Shaoxuan Wang, Shengkai Fang, Shuyi Chen, Sihua Zhou, Stefan Richter, Stephan Ewen, Theodore Vasiloudis, Thomas Weise, Till Rohrmann, Timo Walther, Tzu-Li

(Gordon) Tai, Ufuk Celebi, Vasiliki Kalavri, Volker Markl, Wei Zhong, Weijie Guo, Xiaogang Shi, Xiaowei Jiang, Xingbo Huang, Xingcan Cui, Xintong Song, Yang Wang, Yangze Guo, Yingjie Cao, Yu Li, Yuan Mei, Yun Gao, Yun Tang, Yuxia Luo, Zhijiang Wang, Zhipeng Zhang, Zhu Zhu, Zili Chen (2023); Zhaojing Luo, Beng Chin Ooi, Wei Wang, Meihui Zhang, Qingchao Cai, Shaofeng Cai, Gang Chen, Tien Tuan Anh Dinh, Jinyang Gao, Qian Lin, Shicong Lin, Kee Yuan Ngiam, Gene Yan Ooi, Moaz Reyad, Kian-Lee Tan, Anthony K. H. Tung, Sheng Wang, Yuncheng Wu, Zhongle Xie, Naili Xing, Rulin Xing, Wanqi Xue, Sai Ho Yeung, James Yip, Lingze Zeng, Zhaoqi Zhang, Kaiping Zheng, Lei Zhu, Ji Wang (2024); James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Sebastian Kanthak, Alexander Lloyd, Sergey Melnik, David Mwaura, Sean Quinlan, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, Dale Woodford, David F. Bacon, Shannon Bales, Nico Bruno, Brian F. Cooper, Adam Dickinson, Campbell Fraser, Milind Joshi, Eugene Kogan, Rajesh Rao, David Shue, Marcel van der Holst, Cliff Frey, Damian Reeves, Steve Middlekauff, Mert Akdere, Ben Vandiver, Dan Glick, David Ziegler, Alex Khesin, Dave Weissman, Todd Lipcon, Sean Dorward, Eric Veach (2025).

SIGMOD Contributions Award

For significant contributions to the field of database systems through research funding, education, and professional services. Recipients of the award are the following:

Maria Zemankova (1992)
Gio Wiederhold (1995)
Yahiko Kambayashi (1995)
Jeffrey Ullman (1996)
Avi Silberschatz (1997)
Won Kim (1998)
Raghu Ramakrishnan (1999)
Michael Carey (2000)
Laura Haas (2000)
Daniel Rosenkrantz (2001)
Richard Snodgrass (2002)
Michael Ley (2003)
Surajit Chaudhuri (2004)
Hongjun Lu (2005)
Tamer Özsu (2006)
Hans-Jörg Schek (2007)
Klaus R. Dittrich (2008)
Beng Chin Ooi (2009)
David Lomet (2010)
Gerhard Weikum (2011)
Marianne Winslett (2012)
H.V. Jagadish (2013)
Kyu-Young Whang (2014)
Curtis Dyreson (2015)
Samuel Madden (2016)
Yannis E. Ioannidis (2017)
Z. Meral Özsoyoğlu (2018)
Ahmed Elmagarmid (2019)
Philipe Bonnet (2020)
Juliana Freire (2020)
Stratos Idreos (2020)
Stefan Manegold (2020)
Ioana Manolescu (2020)
Dennis Shasha (2020)
Divesh Srivastava (2021)
Christian S. Jensen (2022)
K. Selcuk Candan (2023)
Sihem Amer-Yahia (2024)
Hector Munoz-Avila & Sylvia Spengler (2025)

SIGMOD Jim Gray Doctoral Dissertation Award

SIGMOD has established the annual SIGMOD Jim Gray Doctoral Dissertation Award to *recognize excellent research by doctoral candidates in the database field.* Recipients of the award are the following:

- **2006** *Winner*: Gerome Miklau. *Honorable Mentions*: Marcelo Arenas and Yanlei Diao
- **2007** *Winner*: Boon Thau Loo. *Honorable Mentions*: Xifeng Yan and Martin Theobald
- **2008** *Winner*: Ariel Fuxman. *Honorable Mentions*: Cong Yu and Nilesh Dalvi
- **2009** *Winner*: Daniel Abadi. *Honorable Mentions*: Bee-Chung Chen and Ashwin Machanavajjhala
- **2010** *Winner:* Christopher Ré. *Honorable Mentions*: Soumyadeb Mitra and Fabian Suchanek
- **2011** *Winner*: Stratos Idreos. *Honorable Mentions*: Todd Green and Karl Schnaitterz
- **2012** *Winner*: Ryan Johnson. *Honorable Mention*: Bogdan Alexe
- **2013** *Winner*: Sudipto Das, *Honorable Mention*: Herodotos Herodotou and Wenchao Zhou
- **2014** *Winners*: Aditya Parameswaran and Andy Pavlo.
- **2015** *Winner*: Alexander Thomson. *Honorable Mentions*: Marina Drosou and Karthik Ramachandra
- **2016** *Winner*: Paris Koutris. *Honorable Mentions*: Pinar Tozun and Alvin Cheung
- **2017** *Winne*r: Peter Bailis. *Honorable Mention*: Immanuel Trummer
- **2018** *Winne*r: Viktor Leis. *Honorable Mention*: Luis Galárraga and Yongjoo Park
- **2019** *Winne*r: Joy Arulraj. *Honorable Mention*: Bas Ketsman
- **2020** *Winner:* Jose Faleiro. *Honorable Mention:* Silu Huang
- **2021** *Winner:* Huanchen Zhang, *Honorable Mentions:* Erfan Zamanian, Maximilian Schleich, and Natacha Crooks
- **2022** *Winner:* Chenggang Wu, *Honorable Mentions:* Pingcheng Ruan and Kexin Rong

- **2023** *Winner:* Supun Nakandala, *Honorable Mentions:* Benjamin Hilprecht and Zongheng Yang
- **2024** *Winner:* Daniel Kang, *Honorable Mentions:* Wei Dong, Jialin Ding, and Yisu Remy Wang

A complete list of all SIGMOD Awards is available at: **https://sigmod.org/sigmod-awards/**

[Last updated: June 1, 2025]

Editor's Notes

Welcome to the June 2025 issue of the ACM SIGMOD Record!

This issue starts with the Database Principles column presenting an article by Atserias and Kolaitis on the topic of consistency of database relations. The authors consider two types of consistency: *local* consistency, in which each pair of relations in the given collection of relations agrees on the values of their shared attributes, and *global* consistency, in which there exists a single "universal" relation such that all the relations in the given collection are its projections. Such a relation is referred to as a *consistency witness* for the relation collection. The article is the first to study in depth the notion of consistency witnesses for relations. The framework of annotated relations and the results presented by the authors cover (among other cases) both standard relational databases and bag databases, which makes the article relevant to both theory and practice of relational databases.

In the Vision column, a contribution by Mohammed and colleagues puts forward the goal of developing effective and efficient assessment procedures for each data-quality dimension in the context of a given data set and use case. Toward achieving this vision, they propose a new perspective on data-quality research, which isolates and studies facets responsible for appropriate assessment procedures across data-quality dimensions. This perspective brings to life a cross-community agenda aiming at integrating technologies for data-quality assessment through the lens of these facets. While the article focuses on structured data, the authors posit that their vision can also be extended to semi-structured and unstructured data.

The Surveys column features an article by Khan and colleagues that provides an overview of synergies between graph data management and graph machine learning. The article focuses both on how graph data management enhances graph machine learning, as well as on how graph machine learning aids in graph data management, with a focus on applications such as query answering over knowledge graphs and data-science tasks. The authors discuss open problems and delineate important directions for research in this space.

The Reminiscences on Influential Papers column, edited by Pınar Tözün, presents contributions by Zoi Kaoudi, Fatemeh Nargesian, and Niv Dayan.

The Advice to Mid-Career Researchers column presents a contribution by Sihem Amer-Yahia, who shares her thoughts and experiences on a number of issues, including mid-career choices and responsibilities, learning from senior and junior colleagues, rethinking time management, working on and moving on from relationships, learning from your failures, and understanding the meaning of your success. The article provides advice on many aspects of the mid-career stage of life, and points out rewards of the journey in your best job in the world.

The DBrainstorming column, whose goal is to discuss new and potentially controversial ideas that might be of interest and benefit to the research community, features an article by Ana Klimovic that considers challenges and opportunities in programming cloud-native applications. The article points out that today's cloud-programming model captures little about the resource requirements and data-access patterns of individual applications. This, in turn, gives rise to a major optimization obstacle in this space. The author calls for rethinking the cloud-programming model by adopting a new paradigm in which users could develop applications that explicitly separate pure-compute functions with I/O

functions, and talks about the current exploration of these ideas in a new Dandelion serverless platform.

The Distinguished Profiles column features an interview with Themis Palpanas, Distinguished Professor of Computer Science at Université Paris Cité, Senior Fellow of the French University Institute (IUF), Head of the Computer Science Department at the Université Paris Cité, and Director of the Data Intelligence Institute of Paris (diiP). In the interview, Themis discusses interactions between time series and data management, comparative advantages of LLMs and alternative technologies in addressing research problems, and his work on entity recognition and data integration. He also shares his perspective on collaborations, setting up and managing the DiiP institute, and on working with students. In the context of his life experience, Themis talks about his experiences in the various places he has been to, his photography and snowboarding hobbies, and foodie secrets from Paris.

The Reports column features two contributions. The first article, by Khan and colleagues, presents outcomes of the LLM+KG workshop that was co-located with VLDB 2024 in Guangzhou, China. The workshop focused on data-management challenges and opportunities arising from effective interactions between LLMs and knowledge graphs. The report outlines perspectives and approaches presented by speakers during the workshop.

The second contribution, by Bikakis and colleagues, reports on the results of responses by experts in the field to the survey conducted by the organizing committee of the International Workshop on Big Data Visual Exploration and Analytics (BigVis) held in 2024. The prespectives gained from the survey responses shed light on challenges, emerging topics, and opportunities stemming from human-data interaction and visual analytics in the AI era.

The issue closes with an Open Forum column, which presents an article by Bhowmick and Srivastava. The authors examine the review-board characteristics of four major data-management conferences across four diversity dimensions over time. The article sets the goal of creating more diverse and balanced review boards, and advocates for the development of tools to support this process.

On behalf of the SIGMOD Record Editorial board, I hope that you enjoy reading the June 2025 issue of the SIGMOD Record!

Your submissions to the SIGMOD Record are welcome via the submission site:
https://mc.manuscriptcentral.com/sigmodrecord

Prior to submission, please read the Editorial Policy on the SIGMOD Record's website:
https://sigmodrecord.org/sigmod-record-editorial-policy/

Rada Chirkova

June 2025

Past SIGMOD Record Editors:

Yanlei Diao (2014-2019)	Ioana Manolescu (2009-2013)	Alexandros Labrinidis (2007–2009)
Mario Nascimento (2005–2007)	Ling Liu (2000–2004)	Michael Franklin (1996–2000)
Jennifer Widom (1995–1996)	Arie Segev (1989–1995)	Margaret H. Dunham (1986–1988)
Jon D. Clark (1984–1985)	Thomas J. Cook (1981–1983)	Douglas S. Kerr (1976-1978)
Randall Rustin (1974-1975)	Daniel O'Connell (1971–1973)	Harrison R. Morse (1969)

Consistency Witnesses for Annotated Relations

Albert Atserias
Universitat Politècnica de Catalunya &
Centre de Recerca Matemàtica
Barcelona, Catalonia, Spain
atserias@cs.upc.edu

Phokion G. Kolaitis
UC Santa Cruz & IBM Research
Santa Cruz, California, USA
kolaitis@ucsc.edu

ABSTRACT

The study of local consistency vs. global consistency of database relations received considerable attention in the early days of relational database theory. In a recent paper, we investigated the notions of local consistency and global consistency for annotated relations, where the annotations come from a positive commutative monoid. One of the differences from the classical case is that the join of two consistent annotated relations need not always be a witness of their consistency. Here, we bring to center stage the notion of a consistency witness function for annotated relations, investigate the properties of consistency witness functions, and provide a new perspective to understanding the interplay between local and global consistency for annotated relations.

1 Introduction

During the past two decades, there has been a growing body of research on annotated databases, i.e., databases in which each fact is annotated with a value from some algebraic structure. This framework generalizes both standard relational databases, where the annotations are 1 (true) and 0 (false), and bag databases, where the annotations are non-negative integers denoting the multiplicity of a fact in the database. Much of the work in this area uses annotations from the universe of some fixed semiring $\mathbb{K} = (K, +, \times, 0, 1)$, where the addition operation $+$ is used to model "alternative" information (e.g., disjunction or existential quantification), while the multiplication operation $\times$ is used to model "joint" information (e.g., conjunction or universal quantification). For this reason, the term *semiring semantics* is often used to refer to the work in this area. Database provenance was the first extensively studied topic in this framework [8, 10, 4]. Subsequent studies focused on conjunctive query containment for annotated databases [7, 12], semiring semantics for first-order logic [6], and evaluation of Datalog programs under semiring semantics [11].

Since the early days of the relational database model, the study of consistency of relations has received significant attention [9, 3, 5]. By definition, a collection of relations $R_1, \ldots, R_m$ is *globally consistent* if there is a relation T such that the projection of T on the attributes of R_i is equal to R_i, for each $i = 1, \ldots, m$. We call such a relation T a *consistency witness for* $R_1, \ldots, R_m$. It is well known that if the collection $R_1, \ldots, R_m$ is globally consistent, then the join $R_1 \bowtie \cdots \bowtie R_m$ is a consistency witness for $R_1, \ldots, R_m$; in fact, it is the largest such consistency witness (see, e.g., [9]). As pointed out in [1], however, the state of affairs is different for bags, since there are two bags that are consistent but their join is not a consistency witness for them; moreover, no largest consistency witness for these bags exists.

In [2], we carried out an investigation of the consistency of annotated relations. Since the definition of consistency of annotated relations involves only the projection operation on relations and since projection is defined using only addition $+$, we considered annotated relations in which the annotations come from a monoid $\mathbb{K} = (K, +, 0)$. The main focus of that investigation was the interplay between local consistency and global consistency, that is, under what conditions a collection of pairwise consistent relations $R_1, \ldots, R_m$ is globally consistent. In particular, we identified a condition on monoids, called the *transportation property*, and showed that a positive monoid $\mathbb{K} = (K, +, 0)$ has the transportation property if and only if every acyclic hypergraph H has the local-to-global consistency property for $\mathbb{K}$-relations, which means that every pairwise consistent collection of $\mathbb{K}$-relations over H is globally consistent. This finding generalizes results about local vs. global consistency for standard relations in [3], as well as results about local vs. global consistency for bags in [1].

In this paper, we bring to front stage the notion of a *consistency witness function* on a positive monoid $\mathbb{K}$, that is to say, a function W that, given two $\mathbb{K}$-relations R and S, returns a $\mathbb{K}$-relation $W(R, S)$ that is a consistency witness for R and S, provided that R and S are consistent $\mathbb{K}$-relations. While the notion of a consistency witness function on $\mathbb{K}$ underlies much of the work in [2], it has not been studied in its own right thus far. Our goal is to make the case that this is a fundamental

notion whose study is well deserved.

After presenting some basic properties of consistency witness functions on $\mathbb{K}$, we introduce the two notions of a c-*join-expression* and a *monotone* c-*join expression* for a consistency witness function on $\mathbb{K}$. These notions extend the notions of join expression and monotone join expressions for the standard join $\bowtie$ operation and for standard relations in [3, 5]. We then establish that the transportation property of a positive monoid $\mathbb{K}$ can be characterized in terms of properties of monotone c-join expressions. Furthermore, we argue that the notion a consistency witness function provides a new perspective to the proofs of the main results in [2]. We elaborate on this new perspective here and, along the way, we discuss methods for defining or constructing consistency witness functions for different types of monoids. Finally, we present some observations concerning the existence of "largest" consistency witness functions for annotated relations. In particular, we point out that a positive monoid being idempotent is a sufficient, but not necessary, condition for the existence of "largest" consistency witness functions for relations annotated with elements from that monoid.

2 Preliminaries

Monoids A *commutative monoid* is a structure $\mathbb{K} = (K, +, 0)$, where $+$ is a binary operation on the universe K of $\mathbb{K}$ that is associative, commutative, and has 0 as its neutral element, i.e., $p + 0 = p = 0 + p$ holds for all $p \in K$. A commutative monoid $\mathbb{K} = (K, +, 0)$ is *positive* if for all elements $p, q \in K$ with $p + q = 0$, we have that $p = 0$ and $q = 0$. From now on, we assume that all commutative monoids considered have at least two elements in their universe.

As an example, the structure $\mathbb{B} = (\{0, 1\}, \vee, 0)$ with disjunction $\vee$ as its operation and 0 (false) as its neutral element is a positive commutative monoid. Other examples of positive commutative monoids include the structure $\mathbb{N} = (Z^{\geq 0}, +, 0)$, and $\mathbb{R}^{\geq 0} = (R^{\geq 0}, +, 0)$, where $Z^{\geq 0}$ is the set of non-negative integers, $R^{\geq 0}$ is the set of non-negative real numbers, and $+$ is the standard addition operation. In contrast, the structure $\mathbb{Z} = (Z, +, 0)$, where Z is the set of integers, is a commutative monoid, but not a positive one. Two examples of positive commutative monoids of different flavor are the structures $\mathbb{T} = (R \cup \{\infty\}, \min, \infty)$ and $\mathbb{V} = ([0, 1], \max, 0)$, where R is the set of real numbers, and $\min$ and $\max$ are the standard minimum and maximum operations. Finally, if A is a set and $\mathcal{P}(A)$ is its powerset, then the structure $\mathbb{P}(A) = (\mathcal{P}(A), \cup, \emptyset)$ is a positive commutative monoid, where $\cup$ is the union operation on sets.

$\mathbb{K}$-relations and their marginals An *attribute* A is a symbol with an associated set $\mathrm{Dom}(A)$ as its *domain*. If X is a finite set of attributes, then we write $\mathrm{Tup}(X)$ for the set of *X-tuples*, i.e., $\mathrm{Tup}(X)$ is the set of functions that take each attribute $A \in X$ to an element of its domain $\mathrm{Dom}(A)$. Note that $\mathrm{Tup}(\emptyset)$ is non-empty as it contains the *empty tuple*, i.e., the unique function with empty domain. If $Y \subseteq X$ is a subset of attributes and t is an X-tuple, then the *projection of t on Y*, denoted by $t[Y]$, is the unique Y-tuple that agrees with t on Y. In particular, $t[\emptyset]$ is the empty tuple.

Let $\mathbb{K} = (K, +, 0)$ be a positive commutative monoid and let X be a finite set of attributes. A *$\mathbb{K}$-relation over X* is a function $R : \mathrm{Tup}(X) \to K$ that assigns a value $R(t)$ in K to every X-tuple t in $\mathrm{Tup}(X)$. We will often write $R(X)$ to indicate that R is a $\mathbb{K}$-relation over X, and we will refer to X as the set of attributes of R. These notions make sense even if X is the empty set of attributes, in which case a $\mathbb{K}$-relation over X is simply a single value from K that is assigned to the empty tuple. Clearly, the $\mathbb{B}$-relations are just the standard relations, while the $\mathbb{N}$-relations are the *bags* or *multisets*, i.e., each tuple has a non-negative integer associated with it that denotes the *multiplicity* of the tuple.

The *support* $\mathrm{Supp}(R)$ of a $\mathbb{K}$-relation $R(X)$ is the set of X-tuples t that are assigned non-zero value, i.e.,

$$\mathrm{Supp}(R) := \{t \in \mathrm{Tup}(X) : R(t) \neq 0\}. \quad (1)$$

We will often write R' to denote $\mathrm{Supp}(R)$. Note that R' is a standard relation over X. A $\mathbb{K}$-relation is *finitely supported* if its support is a finite set. In this paper, all $\mathbb{K}$-relations considered will be finitely supported, and we omit the term; thus, from now on, a $\mathbb{K}$-relation is a finitely supported $\mathbb{K}$-relation. When R' is empty, we say that R is the empty $\mathbb{K}$-relation over X.

If $Y \subseteq X$, then the *marginal $R[Y]$ of R on Y* is the $\mathbb{K}$-relation over Y such that for every Y-tuple t, we have that

$$R[Y](t) := \sum_{\substack{r \in R' : \\ r[Y] = t}} R(r). \quad (2)$$

The value $R[Y](t)$ is the *marginal of R over t*. In what follows and for notational simplicity, we will often write $R(t)$ for the marginal of R over t, instead of $R[Y](t)$. It will be clear from the context (e.g., from the arity of the tuple t) if $R(t)$ is indeed the marginal of R over t (in which case t must be a Y-tuple) or $R(t)$ is the actual value of R on t as a mapping from $\mathrm{Tup}(X)$ to K (in which case t must be an X-tuple). Note that if R is a standard relation (i.e., R is a $\mathbb{B}$-relation), then the marginal $R[Y]$ is the projection of R on Y.

The proof of the next basic proposition follows easily from the definitions.

PROPOSITION 1. *Let $\mathbb{K}$ be a positive commutative monoid and let $R(X)$ be a $\mathbb{K}$-relation. Then the following hold:*

1. *For all $Y \subseteq X$, we have $R'[Y] = R[Y]'$.*
2. *For all $Z \subseteq Y \subseteq X$, we have $R[Y][Z] = R[Z]$.*

If X and Y are sets of attributes, then we write XY as shorthand for the union $X \cup Y$. Accordingly, if x is an X-tuple and y is a Y-tuple such that $x[X \cap Y] = y[X \cap Y]$, then we write xy to denote the XY-tuple that agrees with x on X and on y on Y. We say that x *joins with* y, and that y *joins with* x, to *produce* the tuple xy.

A *schema* is a sequence $X_1, \ldots, X_m$ of sets of attributes. A schema can also be identified with a hypergraph H having $X_1, \ldots, X_m$ as its hyperedges. We will use the terms *schema* and *hypergraph* interchangeably. A *collection of $\mathbb{K}$-relations* over such a schema is a sequence $R_1(X_1), \ldots, R_m(X_m)$ of $\mathbb{K}$-relations so that $R_i(X_i)$ is a $\mathbb{K}$-relation over X_i, for $i = 1, \ldots, m$.

3 Consistency and Consistency Witnesses

Let $\mathbb{K} = (K, +, 0)$ be a positive commutative monoid.

We say that two $\mathbb{K}$-relations $R(X)$ and $S(Y)$ are *consistent* if there is a $\mathbb{K}$-relation $T(XY)$ such that $T[X] = R$ and $T[Y] = S$. Such a $\mathbb{K}$-relation T is called a *consistency witness* for R and S.

A *consistency witness function on* $\mathbb{K}$ is a binary function W that takes as arguments two $\mathbb{K}$-relation $R(X)$ and $S(Y)$, and returns as value a $\mathbb{K}$-relation $W(R, S)$ over XY such that if R and S are consistent $\mathbb{K}$-relations, then $W(R, S)$ is a consistency witness for R and S. For example, the join $\bowtie$ of two standard relations is a consistency witness function on the Boolean monoid $\mathbb{B}$.

We say that a collection $R_1(X_1), \ldots, R_m(X_m)$ of $\mathbb{K}$-relations over a schema $X_1, \ldots, X_m$ is *globally consistent* if there is a $\mathbb{K}$-relation $T(X_1 \ldots X_m)$ such that $T[X_i] = R_i$, for i with $1 \leq i \leq m$. Such a $\mathbb{K}$-relation T is called a *consistency witness* for $R_1, \ldots, R_m$.

It is easy to see that if $R_1(X_1), \ldots, R_m(X_m)$ is a globally consistent collection of $\mathbb{K}$-relations, then these relations are pairwise consistent. Indeed, if T is a consistency witness for $R_1(X_1), \ldots, R_m(X_m)$, then for all i and j with $1 \leq i, j \leq m$, we have that the $\mathbb{K}$-relation $T[X_iX_j]$ is a consistency witness for R_i and R_j, because

$$R_i = T[X_i] = T[X_iX_j][X_i]$$
$$R_j = T[X_j] = T[X_iX_j][X_j]$$

where, in each case, the first equality follows from the definition of global consistency and the second equality follows from Proposition 1.

The converse is known to fail, even for standard relations, i.e., there are standard relations that are pairwise consistent but not globally consistent. The main result by Beeri et al. [3] characterizes the schemas for which the pairwise consistency of a collection of standard relations implies that they are globally consistent. Later on in this paper, we will see how this result extends to $\mathbb{K}$-relations over positive monoids satisfying a condition we call the *transportation property*.

We are interested in obtaining global consistency witnesses by using consistency witnesses for two relations. To this effect, we introduce certain syntactic expressions, which, under some additional hypotheses, will give rise to global consistency witnesses. In what follows, $\bowtie_{\mathrm{c}}$ is a binary function symbol, which will be interpreted by some consistency witness function.

Assume that $X_1, \ldots, X_m$ is a schema.

The collection of c-*join expressions over* $X_1, \ldots, X_m$ is the smallest collection of strings that contains each X_i and has the property that if E_1 and E_2 are in the collection, then also the string $(E_1 \bowtie_{\mathrm{c}} E_2)$ is in the collection.

The collection of *sequential* c-*join expressions over* $X_1, \ldots, X_m$ is the smallest collection of strings that contains each X_i and has the property that if E in the collection and X is one of the X_i's, then also the string $(E \bowtie_{\mathrm{c}} X)$ is in the collection.

A c-join expression over $X_1, \ldots, X_m$ is called *read-once* if each X_i appears exactly once in the expression. We write $E[X_1, \ldots, X_m]$ to denote the read-once sequential c-join-expression on $X_1, \ldots, X_m$ where the X_i appear in the indicated order; in the sequel, we refer to $E[X_1, \ldots, X_m]$ as the *read-once sequential* c-*join-expression associated with the ordering* $X_1 \ldots, X_m$. In symbols, we have that $E[X_1, \ldots, X_m]$ is the c-join expression

$$(\cdots((X_1 \bowtie_{\mathrm{c}} X_2) \bowtie_{\mathrm{c}} X_3) \bowtie_{\mathrm{c}} \cdots \bowtie_{\mathrm{c}} X_m).$$

Clearly, the string $((X_1 \bowtie_{\mathrm{c}} X_2) \bowtie_{\mathrm{c}} X_3)$ is a sequential c-join-expression, while the string

$$((X_1 \bowtie_{\mathrm{c}} X_2) \bowtie_{\mathrm{c}} (X_3 \bowtie_{\mathrm{c}} X_4))$$

is a c-join expression, but not a sequential one. Furthermore, both these strings are read-once c-join expressions, while $((X_1 \bowtie_{\mathrm{c}} X_2) \bowtie_{\mathrm{c}} (X_3 \bowtie_{\mathrm{c}} X_1))$ is not. From now on we drop the outermost parentheses.

The notion of a c-join expression is a syntactic one. We will now assign semantics to c-join expressions.

Let $X_1, \ldots, X_m$ be a schema and let E be a c-join-expression over $X_1, \ldots, X_m$. If W is a consistency witness function on $\mathbb{K}$ and $R_1(X_1), \ldots, R_m(X_m)$ is a collection of $\mathbb{K}$-relations, we write $E(W, R_1, \ldots, R_m)$ to denote the $\mathbb{K}$-relation over $X_1 \cdots X_m$ obtained by evaluating E when $\bowtie_{\mathrm{c}}$ is interpreted by W and each X_i is interpreted by R_i for $i = 1, \ldots, m$.

We say that E is *monotone with respect to W and $R_1, \ldots, R_m$* if for every sub-expression $E_1 \bowtie_{\mathrm{c}} E_2$ of E, we have that the $\mathbb{K}$-relations $E_1(W, R_1, \ldots, R_m)$ and $E_2(W, R_1, \ldots, R_m)$ are consistent.

According to the next proposition, monotone c-join-expressions give rise to global consistency witnesses.

PROPOSITION 2. *Let E be a* c*-join expression over $X_1, \ldots, X_m$, let W be a consistency witness function on $\mathbb{K}$, and let $R_1(X_1), \ldots, R_m(X_m)$ be $\mathbb{K}$-relations. If E is monotone with respect to W and $R_1, \ldots, R_m$, and every X_i occurs in E, then $E(W, R_1, \ldots, R_m)$ is a global consistency witness for the $\mathbb{K}$-relations $R_1, \ldots, R_m$.*

This proposition is proved by induction on the construction of c-join expressions.

The base case is trivial, since in this case E is X_i for some i with $1 \leq i \leq n$, hence $E(W, R_i) = R_i$, which is a consistency witness for R_i.

For the inductive step, assume that E is $E_1 \bowtie_{\mathrm{c}} E_2$, where E_1 and E_2 are c-join expressions for which the inductive hypothesis holds. To simplify the notation, let us put $\mathbf{R} = (R_1, \ldots, R_m)$; furthermore, we put $\mathbf{R}_1 = (R_i : i \in I_1)$ and $\mathbf{R}_2 = (R_i : i \in I_2)$, where I_1 and I_2 are the sets of indices i such that X_i occurs in E_1 and in E_2, respectively. In this case, we have that $E(W, \mathbf{R}) = W(E_1(W, \mathbf{R}_1), E_2(W, \mathbf{R}_2))$.

Since E is monotone with respect to W and $\mathbf{R}$, we have that the $\mathbb{K}$-relations $E_1(W, \mathbf{R}_1)$ and $E_2(W, \mathbf{R}_2)$ are consistent, hence $W(E_1(W, \mathbf{R}_1), E_2(W, \mathbf{R}_2))$ is a consistency witness for $E_1(W, \mathbf{R}_1)$ and $E_2(W, \mathbf{R}_2)$. We must show that $W(E_1(W, \mathbf{R}_1), E_2(W, \mathbf{R}_2))[X_i] = R_i$ holds, for every i such that X_i occurs in E. Consider such an X_i. Since X_i occurs in E, it must occur in at least one of E_1 and E_2. Let's assume that X_i occurs in E_1; the case in which it occurs in E_2 is entirely similar. If Y is the set of attributes of $E_1(W, \mathbf{R}_1)$, then $X_i \subseteq Y$. Furthermore, the property of an expression being monotone with respect to a witness function and a collection of relations is inherited by its subexpressions, so E_1 is monotone with respect to W and $\mathbf{R}_1$. By induction hypothesis, $E_1(W, \mathbf{R}_1)$ is a global consistency witness of all relations R_j occurring in it, hence

$$E_1(W, \mathbf{R}_1)[X_i] = R_i. \tag{3}$$

Also, since $W(E_1(W, \mathbf{R}_1), E_2(W, \mathbf{R}_2))$ is a consistency witness for $E_1(W, \mathbf{R}_1)$ and $E_2(W, \mathbf{R}_2)$, we have that

$$W(E_1(W, \mathbf{R}_1), E_2(W, \mathbf{R}_2))[Y] = E_1(W, \mathbf{R}_1). \tag{4}$$

By putting everything together, we have that

$$\begin{aligned} &W(E_1(W, \mathbf{R}_1), E_2(W, \mathbf{R}_2))[X_i] \\ &= W(E_1(W, \mathbf{R}_1), E_2(W, \mathbf{R}_2))[Y][X_i] \\ &= E_1(W, \mathbf{R}_1)[X_i] \\ &= R_i, \end{aligned}$$

where in the first equality we used Proposition 1 and the fact that $X_i \subseteq Y$, in the second we used (4), and the third is (3). This completes the proof of Proposition 2.

4 The Transportation Property

We consider several different properties of monoids and establish that they are equivalent to each other.

Let $\mathbb{K} = (K, +, 0)$ be a positive commutative monoid.

If m and n are positive integers, we say that $\mathbb{K}$ has the $m \times n$ *transportation property* if for every column m-vector $b = (b_1, \ldots, b_m) \in K^m$ with entries in K and every row n-vector $c = (c_1, \ldots, c_n) \in K^n$ with entries in K such that $b_1 + \cdots + b_m = c_1 + \cdots + c_n$ holds, there is an $m \times n$ matrix $D = (d_{ij} : i \in [m], j \in [n]) \in K^{m \times n}$ with entries in K whose rows sum to b and whose columns sum to c, i.e., $d_{i1} + \cdots + d_{im} = b_i$ for all $i \in [m]$ and $d_{1j} + \cdots + d_{mj} = c_j$ for all $j \in [n]$.

We say that $\mathbb{K}$ has the *transportation property* if $\mathbb{K}$ has the $m \times n$ transportation property for every pair (m, n) of positive integers.

We now consider a number of properties of monoids that involve $\mathbb{K}$-relations.

Two $\mathbb{K}$-relations $R(X)$ and $S(Y)$ are *inner consistent* if $R[X \cap Y] = S[X \cap Y]$. Using Proposition 1, it is easy to verify that if R and S are consistent $\mathbb{K}$-relations, then they are also inner consistent. The converse, however, is not true for all positive commutative monoids. We single out the ones for which inner consistency implies consistency (consequently, for such monoids, these two notions are equivalent).

We say that $\mathbb{K}$ has the *inner consistency property* if whenever two $\mathbb{K}$-relations are inner consistent, then they are also consistent.

We say that a schema $X_1, \ldots, X_m$ has the *local-to-global consistency property for $\mathbb{K}$-relations* if every collection $R_1(X_1), \ldots, R_m(X_m)$ of pairwise consistent $\mathbb{K}$-relations is also globally consistent.

Let E be a c-join-expression over $X_1, \ldots, X_m$. We say that E is *monotone on* $\mathbb{K}$ if E is monotone with respect to every consistency witness function W on $\mathbb{K}$ and every collection $R_1(X_1), \ldots, R_m(X_m)$ of pairwise consistent $\mathbb{K}$-relations.

Finally, we say that a schema $X_1, \ldots, X_m$ *admits a monotone* c*-join expression on* $\mathbb{K}$ if there is a c-join-expression E over $X_1, \ldots, X_m$ that is monotone on $\mathbb{K}$ and, furthermore, every X_i occurs in E.

THEOREM 1. *The following statements are equivalent for a positive monoid $\mathbb{K}$:*

1. *$\mathbb{K}$ has the 2×2 transportation property.*
2. *$\mathbb{K}$ has the transportation property.*
3. *$\mathbb{K}$ has the inner consistency property.*
4. *Every acyclic hypergraph admits a monotone read-once sequential* c*-join-expression on $\mathbb{K}$.*
5. *Every acyclic hypergraph admits a monotone read-once* c*-join-expression on $\mathbb{K}$.*

6. *Every acyclic hypergraph admits a monotone* c*-join-expression on* $\mathbb{K}$.
7. *Every acyclic hypergraph has the local-to-global consistency property for* $\mathbb{K}$*-relations.*

The proofs of the implications $(1) \Rightarrow (2)$ and $(2) \Rightarrow (3)$ are given in [2]. A new perspective on these proofs will be presented in Section 6. Here, we sketch the proofs of the remaining implications in a round-robin fashion.

We begin with the implication $(3) \Rightarrow (4)$. As shown in Beeri et al. [3], if H is an acyclic hypergraph, then H has the *running intersection* property, which means that there is an ordering $X_1, \ldots, X_m$ of the hyperedges of H so that for every $j \leq m$, there is some $i \leq j-1$ such that $(X_1 \cup \cdots \cup X_{j-1}) \cap X_j \subseteq X_i$. Let $E[X_1, \ldots, X_m]$ be the read-once sequential c-join-expression associated with this ordering, i.e., $E[X_1, \ldots, X_m]$ is

$$(\cdots((X_1 \bowtie_{\mathrm{c}} X_2) \bowtie_{\mathrm{c}} X_3) \bowtie_{\mathrm{c}} \cdots \bowtie_{\mathrm{c}} X_m).$$

Using the inner consistency property of $\mathbb{K}$, it is not hard to show that $E[X_1, \ldots, X_m]$ is monotone on $\mathbb{K}$. The implications $(4) \Rightarrow (5)$ and $(5) \Rightarrow (6)$ are trivial. The implication $(6) \Rightarrow (7)$ uses Proposition 2 and, of course, the definitions.

Finally, we prove $(7) \Rightarrow (1)$. We are given a 2×2 instance of the transportation problem on $\mathbb{K}$: four elements $b_1, b_2, c_1, c_2 \in K$ such that $b_1 + b_2 = c_1 + c_2$. Consider the following three $\mathbb{K}$-relations where $e = b_1 + b_2 = c_1 + c_2$:

AB	BC	CD
$1\,0 : b_1$	$0\,0 : e$	$0\,1 : c_1$
$2\,0 : b_2$	$1\,1 : e$	$0\,2 : c_2$
$1\,1 : c_1$		$1\,1 : b_1$
$2\,1 : c_2$		$1\,2 : b_2$

It is easy to see that these are pairwise consistent, and the schema is acyclic as it is the path of length three. By (7) the three $\mathbb{K}$-relations are also globally consistent. Let $W(ABCD)$ be a witness of global consistency. Setting $d_{ij} = W(i00j)$ or $d_{ij} = W(j11i)$ we get a solution to the 2×2 instance, which completes the proof.

Beeri et al. showed that a hypergraph H is acyclic if and only if H has the local-to-global consistency property for standard relations (i.e., for $\mathbb{B}$-relations, where $\mathbb{B}$ is the Boolean monoid). In [2], we showed that if $\mathbb{K}$ is an arbitrary positive monoid and H is a hypergraph that has the local-to-global consistency property for $\mathbb{K}$-relations, then H must be acyclic. We also showed that there are positive commutative monoids $\mathbb{K}$ and acyclic schemas H that do *not* have the local-to-global consistency property for $\mathbb{K}$-relations. Thus, acyclicity is a necessary, but not sufficient, condition for H to have the local-to-global consistency property for $\mathbb{K}$-relations. Theorem 1, however, implies that acyclicity is both necessary and sufficient, provided $\mathbb{K}$ has the transportation property. Thus, we have the following generalization of the main result in Beeri et al. [3].

THEOREM 2. *Assume that* $\mathbb{K}$ *is a positive commutative monoid that has the transportation property. For every hypergraph* H*, the following statements are equivalent:*

1. H *is acyclic.*
2. H *admits an ordering* $X_1, \ldots, X_m$ *of its hyperedges so that the sequential* c*-join expression associated with* $X_1, \ldots, X_m$ *is monotone on* $\mathbb{K}$.
3. H *admits a monotone* c*-join-expression on* $\mathbb{K}$.
4. H *has the local-to-global consistency property for* $\mathbb{K}$*-relations.*

Naturally, in Theorem 2 we can also add as equivalent statements that H admits a sequential monotone c-join-expression on $\mathbb{K}$, as well as a read-once sequential monotone c-join-expression on $\mathbb{K}$. Recall that this last condition is equivalent to the statement that there is an ordering $X_1, \ldots, X_m$ of the hyperedges of H so that the sequential c-join expression $E[X_1, \ldots, X_m]$ associated with $X_1, \ldots, X_m$ is monotone on $\mathbb{K}$.

5 Defining Consistency Witnesses

By definition, every consistency witness function for a positive commutative monoid $\mathbb{K}$ produces a consistency witness $W = W(R, S)$, given two consistent $\mathbb{K}$-relations $R(X)$ and $S(Y)$. But how can such a consistency witness function be defined? Are there general ways of constructing a consistency witness function?

For several specific monoids of interest, the consistency witness can be found via an explicit expression or via a procedural method. For example, for the Boolean monoid $\mathbb{B}$, the standard join $R \bowtie S$ of standard relations is an explicit consistency witness function. More generally, if $\mathbb{K} = (K, \vee, 0)$ is the join semilattice of a bounded distributive lattice $(K, \vee, \wedge, 0, 1)$ (the same way the Boolean monoid $\mathbb{B}$ is the join semilattice of the 2-element Boolean algebra), then setting

$$W(t) = R(t[X]) \wedge S(t[Y]) \tag{5}$$

for every XY-tuple t gives an explicit expression that defines a consistency witness function for every two consistent $\mathbb{K}$-relations $R(X)$ and $S(Y)$.

Similarly, if $\mathbb{K} = (K, +, 0)$ is the additive monoid of a semifield $(K, +, \times, /, 0, 1)$, (the same way the positive monoid $\mathbb{R}^{\geq 0}$ of non-negative reals with addition is the additive monoid of the semifield of non-negative real numbers with addition and multiplication), then an explicit expression for a consistency witness is given by

setting

$$W(t) = R(t[X]) \times S(t[Y]) / D(t) \tag{6}$$

where $D(t) = R(t[X \cap Y]) = S(t[X \cap Y])$ with the convention that $0/0 = 0$. Note that the equality in the definition of $D(t)$ follows from the assumption that R and S are consistent; indeed, if U witnesses their consistency, then

$$R[X \cap Y] = U[X][X \cap Y] = U[Y][X \cap Y] = S[X \cap Y],$$

where the middle equation follows from Proposition 1.

The expressions in (5) and (6) are called respectively the *standard join* of the distributive lattice, which is denoted by $R \bowtie_{\mathbb{K}} S$, and the *Vorobe'v join* of the semifield, which is denoted by $R \bowtie_{\mathrm{V}\mathbb{K}} S$.

When it comes to the bag monoid $\mathbb{N} = (N, +, 0)$, it turns out that the standard join of bags is *not* a valid consistent witness function. For example, the two bags $R(X) = \{a{:}1, b{:}1\}$ and $S(Y) = \{c{:}1, d{:}1\}$ are consistent via the witness $\{ac{:}1, bd{:}1\}$ or $\{ad{:}1, bc{:}1\}$, but their bag join is the bag $\{ac{:}1, ad{:}1, bc{:}1, bd{:}1\}$, which projects to $\{a{:}2, b{:}2\}$ on X and to $\{c{:}2, d{:}2\}$ on Y, thus it is not a witness of their consistency. Nonetheless, the bag monoid does admit an explicit consistency witness function, which can be defined via a procedure called the Northwest Corner Method. As explained in [2], the inspiration for this procedure came from linear programming, simplifying an earlier method from [1]. In Section 7 we provide an alternative perspective to it.

We refer the reader to Section 5 of [2] for an ample discussion of specific monoids and classes of monoids for which a consistency witness can be explicitly defined by an expression or by a procedural method, such as the Northwest Corner Method.

In the next section, we discuss a more general problem, which is implicit in the validity of the implications (1) $\Rightarrow$ (2) $\Rightarrow$ (3) of Theorem 1. The problem can be stated as follows: How can the transportation property alone be used to construct consistency witnesses in full generality? First we discuss how a direct interpretation of the proof of the implication (2) $\Rightarrow$ (3) in Theorem 1 gives a way to construct consistency witnesses by solving explicit but typically large systems of equations over the monoid. Then we argue that the proof of the implication (1) $\Rightarrow$ (2) in Theorem 1 indeed gives a way to construct witnesses from just solving 2×2 systems.

This is a rather remarkable phenomenon that enables the construction of consistency witnesses by repeatedly solving many but tiny 2×2 systems of equations over the monoid. This phenomenon is akin to the fact that the standard join of standard relations can be computed very efficiently (in terms of the output size) by scanning the pairs of tuples in the two relations in a carefully chosen order. As we will see, in the general case of positive commutative monoids with the transportation property, it suffices to scan not pairs of tuples (i.e., 1×1 systems) but *pairs of pairs of tuples* (i.e., 2×2 systems), also in some suitable order.

We begin our discussion by recalling the aforementioned standard and efficient method for computing joins of standard relations.

6 From 2×2 Systems to Witnesses

For relational databases, the Sort-Merge Join algorithm is a well-known method to compute the join of two relations $R(X)$ and $S(Y)$; e.g., see Section 12.5.2 in [13]. The algorithm works as follows.

First sort the tuples in R and S in the two relations lexicographically by the entries of the tuples on the common attributes $Z = X \cap Y$, i.e., sort all tuples $r \in R$ by $r[Z]$ and sort all tuples $s \in S$ by $s[Z]$. Then, scan the two sorted lists in parallel to find a tuple $t \in \mathrm{Tup}(Z)$ on the common attributes that appears in both lists. For each such t found, scan all pairs of tuples $r \in R$ and $s \in S$ such that $r[Z] = t$ and $s[Z] = t$, produce the join tuple rs in the output $W(XY)$, and proceed to the next common t in the sorted lists. Since the join of two consistent standard relations is a witness of their consistency, this algorithm computes a consistency witness function for the Boolean monoid $\mathbb{B}$.

When the positive monoid $\mathbb{K}$ has the transportation property, there is a natural analogue of the Sort-Merge Join algorithm that produces consistency witnesses for consistent $\mathbb{K}$-relations $R(X)$ and $S(Y)$. Again, first sort all tuples $r \in R'$ and all tuples $s \in S'$ in the supports R' and S' of R and S lexicographically by the entries of the tuples on the common attributes $Z = X \cap Y$. Then, scan the sorted lists to find the first tuple $t \in \mathrm{Tup}(Z)$ that appears in both lists. For such t, form a system of equations over $\mathbb{K}$. For each $r \in R'$ and $s \in S'$ such that $r[Z] = s[Z] = t$, the system has one variable $x_{r,s;t}$. The system has equations

$$\sum_{\substack{s \in S': \\ s[Z]=t}} x_{r,s;t} = b_r \quad \text{and} \quad \sum_{\substack{r \in R': \\ r[Z]=t}} x_{r,s;t} = c_s$$

for each $r \in R'$ with $r[Z] = t$ and $b_r = R(r)$ in the first equation, and each $s \in S'$ with $s[Z] = t$ and $c_s = S(s)$ in the second equation.

Now note that by the assumption that $R(X)$ and $S(Y)$ are consistent, we have $\sum_r b_r = \sum_s c_s$. By the transportation property of $\mathbb{K}$, the system has a solution in $\mathbb{K}$, say by setting $x_{r,s;t}$ to $a_{r,s;t}$. Finally, use this solution to produce the annotated tuple $rs{:}a_{r,s;t}$ in the output $W(XY)$ for each considered r and s, and proceed to the next common t in the sorted lists. The fact that the resulting $\mathbb{K}$-relation $W(XY)$ is a consistency witness for R and S is an immediate consequence of the defi-

nitions and the way the system of equations was set up. This construction is also what goes behind the scenes in the proof of the implication (2) $\Rightarrow$ (3) in Theorem 1. We refer the reader to [2] for more details on this proof.

An important point about the Sort-Merge Join algorithm of the previous paragraph is that it involves solving systems of equations of many different sizes, and often very big ones. Concretely, if for a tuple t that appears in both lists we have m_t tuples $r \in R'$ such $r[Z] = t$ and n_t tuples $s \in S'$ such that $s[Z] = t$, then the system associated to tuple t has $m_t \times n_t$ variables and $m_t + n_t$ equations. Since m_t and n_t could in general be quite big, solving each such system for each tuple t individually could be computationally expensive. This should be compared with the explicit and usually efficiently computable expressions of Equation (5) for the standard join of a distributive lattice, and Equation (6) for the Vorobe'v join of a semifield. In contrast to these explicit expressions, if all we know about the monoid is that it has the transportation property, then no such explicit expression may be available; thus, it looks like we are stuck with the daunting task of solving potentially huge $m_t \times n_t$ systems of equations for each t.

Or are we?

Interestingly, the implication (1) $\Rightarrow$ (2) in Theorem 1 asserts that the 2×2 transportation property *alone* already implies the $m \times n$ transportation property for every positive integers m and n. At least in principle, this means that in order to solve the $m_t \times n_t$ systems of each t it should suffice to solve perhaps many but tiny 2×2 systems. In the rest of this section we explain how the proof of the implication (1) $\Rightarrow$ (2) in Theorem 1 can be leveraged to reduce the task for solving the $m_t \times n_t$ systems within the context of the Sort-Merge Join algorithm to that of solving many but tiny 2×2 systems.

To discuss this, let us first examine one possible implementation of the inner loop in the Sort-Merge Join algorithm for standard relations. The method we suggest below is almost certainly not what would be implemented in practice because, for practical implementations, iterative methods are preferred over recursive ones. However, it is conceptually useful to explain the method as a recursive algorithm to see how it generalizes to the case of $\mathbb{K}$-relations over monoids that have the transportation property.

Within the Sort-Merge Join algorithm for standard relations, let's say we are in the situation where we have detected a tuple $t \in \mathrm{Tup}(Z)$ that appears in both sorted lists of the tuples of $R(X)$ and $S(Y)$. The subroutine that we are about to describe produces all join-tuples rs for $r \in R$ and $s \in S$ such that $r[Z] = s[Z] = t$.

Let the sorted lists of such tuples be $r_1, \ldots, r_m$ and $s_1, \ldots, s_n$, respectively. If $m = 1$, then we output the join tuples $r_1 s_j$ for $j = 1, \ldots, n$ and we are done. Symmetrically, if $n = 1$, then we output $r_i s_1$ for $i = 1, \ldots, m$ and we are done again. Suppose then that $m \geq 2$ and $n \geq 2$. If $m > n$, then we split the problem into a base case with the singleton list r_m and a recursive case with the reduced list $r_1, \ldots, r_{m-1}$. In both cases the other list remains $s_1, \ldots, s_n$. Symmetrically, if $m < n$, then we split the problem into a base case with the singleton list s_n, and a recursive case with the reduced list $s_1, \ldots, s_{n-1}$. Again, in both cases the other list remains $r_1, \ldots, r_m$. In case $m = n$, we just break ties arbitrarily and go with one of the two. Since $m \geq 2$ and $n \geq 2$, the recursive calls made in this subroutine call always make progress in reducing the sizes of the lists and we end up producing all pairs $r_i s_j$, as required.

What we need to answer now is why we cannot just do the same for the variant of the Sort-Merge Join algorithm for consistent $\mathbb{K}$-relations. The base cases $m = 1$ and $n = 1$ can certainly be handled the same way, using the annotations $a_{r_1,s_j;t} = S(s_j)$ for the tuples $r_1 s_j$ in the case $m = 1$, and the annotations $a_{r_i,s_1;t} = R(r_i)$ for the tuples $r_i s_1$ in the case $n = 1$. The problem is with the subroutine call in the inductive case $m \geq 2$ and $n \geq 2$: the recursive call with the reduced list does not interact at all with the call with the singleton list, so it is hard to believe that the two calls will magically produce a solution $a_{r_i,s_j;t}$ that satisfies the equations of the consistency requirement. These equations impose global conditions that involve the full lists $r_1, \ldots, r_m$ and $s_1, \ldots, s_n$; therefore, they require some kind of coordination between calls. It is here where it is useful to upgrade the kind of processing that the algorithm does from handling pairs of tuples to handling pairs of pairs of tuples (i.e., 2×2 systems). Let us see how to do this.

As a reminder, it is useful to keep in mind the following graphical representation of the system of equations that we need to satisfy:

$$\begin{array}{ccccccc} x_{1,1} & + & \cdots & + & x_{1,n} & = & b_1 \\ + & & & & + & & \\ \vdots & & \ddots & & \vdots & & \\ + & & & & + & & \\ x_{m,1} & + & \cdots & + & x_{m,n} & = & b_n \\ \| & & & & \| & & \\ c_1 & & & & c_n & & \end{array} \tag{7}$$

where for simplicity we wrote $x_{i,j}$ instead of $x_{r_i,s_j;t}$ and b_i and c_j instead of $R(r_i)$ and $S(s_j)$.

The solution to the problem of non-interacting calls can be discovered by examining how we would manually handle the next limiting cases after the base cases. Let's say $m = 2$, so the first list of tuples is r_1, r_2 and the column vector in the right-hand side of the system (7) is b_1, b_2, and the second list of tuples is $s_1, \ldots, s_n$ with $n \geq 2$. If we were able to solve any 2×2 instance of the transportation problem, then we could split the problem

of solving the system (7) in the special case $m = 2$ as follows. First we solve the 2×2 system given by the equations

$$\begin{array}{ccccc} y_1 & + & x_{1,n} & = & b_1 \\ + & & + & & \\ y_2 & + & x_{2,n} & = & b_2 \\ || & & || & & \\ c & & c_n & & \end{array}$$

where y_1, y_2 are two new variables and $c = c_1 + \cdots + c_{n-1}$. Observe that $c + c_n = b_1 + b_2$, as required by the transportation property. Once this is solved, we go on recursively to solve the $2 \times (n-1)$ system given by the equations

$$\begin{array}{ccccccc} x_{1,1} & + & \cdots & + & x_{1,n-1} & = & y_1 \\ + & & & & + & & \\ x_{2,1} & + & \cdots & + & x_{2,n-1} & = & y_2 \\ || & & & & || & & \\ c_1 & & & & c_{n-1} & & \end{array}$$

Observe that $c_1 + \cdots + c_{n-1} = c = y_1 + y_2$, as required by the transportation property. Note also how the part y_1, y_2 of the solution to the first system is *used to define* the right-hand side of the second system, so the two calls of the subroutine *now do interact*. A simple inspection shows that the concatenation of the solutions of the two systems gives a solution to the global $m \times n$ system (7) in the special case $m = 2$.

This analysis takes care of the case $m = 2$ and $n \geq 2$. To take care of the case $m \geq 2$ and $n = 2$, we proceed symmetrically exchanging rows and columns. Finally, for the case $m \geq 3$ and $n \geq 3$, we can use the cases $2 \times n$ and $m \times 2$ that we just discussed as base cases. We split an $m \times n$ system as in (7) into a $2 \times n$ system and an $(m-1) \times n$ system if $m > n$, or into an $m \times 2$ system and an $m \times (n-1)$ system if $m < n$, breaking ties arbitrarily if $m = n$. This completes the description and the analysis of the recursive algorithm. The inductive argument that proves its correctness is also what goes behind the scenes in the proof of the implication (1) $\Rightarrow$ (2) in Theorem 1, as presented in [2].

The bottom line of this section is that the Sort-Merge Join algorithm for computing joins of standard relations, and hence consistency witnesses of standard relations, nicely generalizes to an algorithm for computing consistency witnesses of two given consistent $\mathbb{K}$-relations from just knowing how to solve many but explicit and tiny 2×2 instances of the transportation problem.

7 Solving 2×2 Systems in Specific Cases

In view of the analysis of the previous section, it is now natural to revisit the question of solving 2×2 systems for specific monoids. In this section, we revisit the standard join and the Vorobe'v join in Section 5 from the perspective of 2×2 systems. We also give an explicit solution to the 2×2 systems for monoids for which such systems are solvable using the Northwest Corner Method, also mentioned in Section 5. As stated there, the most natural example of this last case is the bag monoid $\mathbb{N}$. By unfolding the recursive algorithm of the previous section, the explicit solution we give in this section gives an alternative and computationally more explicit definition of the Northwest Corner Method, as compared to how it was presented in [2].

Let $\mathbb{K} = (K, +, 0)$ be a positive commutative monoid. We are given b_1, b_2, c_1, c_2 such that $b_1 + b_2 = c_1 + c_2$. We want to solve the following system:

$$\begin{array}{ccccc} x_{11} & + & x_{12} & = & b_1 \\ + & & + & & \\ x_{21} & + & x_{22} & = & b_2 \\ || & & || & & \\ c_1 & & c_2 & & \end{array}$$

We may assume that all b_1, b_2, c_1, c_2 are different from 0 as otherwise we can set both variables of the corresponding row or column equation to 0 and reduce the system to a single trivially satisfiable equation.

Let $e = b_1 + b_2 = c_1 + c_2$ and note $e \neq 0$ by positivity.

If $\mathbb{K} = (K, \vee, 0)$ is the join semilattice of a bounded distributive lattice $(K, \vee, \wedge, 0, 1)$, then setting $x_{ij} = b_i \wedge c_j$ for $i, j = 1, 2$ gives a solution. Indeed, $x_{i1} \vee x_{i2} = (b_i \wedge c_1) \vee (b_i \wedge c_2) = b_i \wedge (c_1 \vee c_2) = b_i \wedge e = b_i \wedge (b_1 \vee b_2) = b_i$. An entirely symmetric argument gives $x_{1j} \vee x_{2j} = c_j$. Examples include the Boolean monoid, the power set monoid, and many others.

If $\mathbb{K} = (K, +, 0)$ is the additive monoid of a semifield $(K, +, \times, /, 0, 1)$, then setting $x_{ij} = (b_i \times c_j)/e$ for $i, j = 1, 2$ gives a solution. Indeed, $x_{i1} + x_{i2} = (b_i \times c_1)/e + (b_i \times c_2)/e = b_i \times (c_1 + c_2)/e = b_i$. Similarly, $x_{1j} + x_{2j} = c_j$. Examples include the nonnegative reals with addition, tropical monoids such as $(R \cup \{-\infty\}, \max, -\infty)$ and many others.

Finally we come to the bag monoid $\mathbb{N} = (Z^{\geq 0}, +, 0)$ and those positive monoids whose instances of the transportation problem can be solved by the Northwest Corner Method. We need some preliminary definitions.

Every positive commutative monoid $\mathbb{K} = (K, +, 0)$ is *canonically preordered* by the binary relation $x \leq y$ defined to hold between two elements $x, y \in K$ if there exists an element $z \in K$ such that $x + z = y$. If for any every two elements $x, y \in K$ we have $x \leq y$ or $y \leq x$ (or both), then we say that this preorder is *total* and that the monoid is *totally canonically preordered*. In such a case, the operation $\min(x, y)$, which returns x if $x \leq y$ and y otherwise, satisfies the inequalities $\min(x, y) \leq x$ and $\min(x, y) \leq y$. Similarly, the operation $\max(x, y)$, which returns y if $x \leq y$ and x otherwise, satisfies the inequalities $x \leq \max(x, y)$ and $y \leq \max(x, y)$. We say that $\mathbb{K}$ is *weakly cancellative* if for every x, y, z, we have

that $x+y=x+z$ implies that $y=z$ or $y=0$ or $z=0$. When a monoid is weakly cancellative, it is natural to define an operation $x \dot{-} y$ on pairs x, y. Concretely, if $x \not\leq y$, then we set $x \dot{-} y = 0$, and if $x \leq y$ via $x+z=y$, then we set $x \dot{-} y = z$ if $x \neq y$ and $x \dot{-} y = 0$ if $x = y$. By weak cancellativity, $x \dot{-} y$ is well defined because if both $x+z=y$ and $x+z'=y$ hold, then $x+z=x+z'$ so by weak cancellativity we have $z = z'$, or $z = 0$ in which case $x = y$, or $z' = 0$ in which case again $x = y$. This operation has the property that if $x \leq y$, then $x + (y \dot{-} x) = y$.

Suppose now that $\mathbb{K}$ is totally canonically preordered and weakly cancellative. The typical example is the bag monoid $\mathbb{N}$, for which $\min(x, y)$ and $\max(x, y)$ are the minimum and the maximum operations, and $\dot{-}$ is the subtraction operation truncated to 0. In this case, a solution is given by the Northwest Corner Method, which in the 2×2 case reduces to the following explicit assignment (recall that $e = b_1 + b_2 = c_1 + c_2$ and b_1, b_2, c_1, c_2 are different from 0):

$$\begin{array}{lcl} x_{11} & = & \min(b_1, c_1) \\ x_{12} & = & b_1 \dot{-} x_{11} \\ x_{21} & = & c_1 \dot{-} x_{11} \\ x_{22} & = & e \dot{-} \max(b_1, c_1) \end{array}$$

To see that this system satisfies the 2×2 system first observe that $x_{11} \leq b_1$ and $x_{11} \leq c_1$, so $x_{11} + x_{12} = x_{11} + (b_1 \dot{-} x_{11}) = b_1$ and $x_{11} + x_{21} = x_{11} + (c_1 \dot{-} x_{11}) = c_1$. This already shows that half of the equations of the 2×2 system are satisfied. For the remaining two equations, first we claim that

$$\begin{array}{ll} b_2 = (c_1 \dot{-} b_1) + c_2 & \text{if } b_1 \leq c_1 \\ c_2 = (b_1 \dot{-} c_1) + b_2 & \text{if } b_1 \not\leq c_1. \end{array} \tag{8}$$

Indeed, if $b_1 \leq c_1$ then $c_1 = b_1 + (c_1 \dot{-} b_1)$, so we have $b_1 + b_2 = c_1 + c_2 = b_1 + (c_1 \dot{-} b_1) + c_2$. The first equality in (8) then follows from weak cancellativity because $b_2 \neq 0$ and $c_2 \neq 0$, and therefore also $(c_1 \dot{-} b_1) + c_2 \neq 0$ by positivity. Similarly, if $b_1 \not\leq c_1$, then we have $c_1 \leq b_1$ because the preorder is total, so $b_1 = c_1 + (b_1 \dot{-} c_1)$ and we have $c_1 + c_2 = b_1 + b_2 = c_1 + (b_1 \dot{-} c_1) + b_2$. The second equality in (8) follows then again by weak cancellativity and positivity. Now we use (8) to show, by cases, that the remaining two equations of the 2×2 system are satisfied.

If $b_1 \leq c_1$, then $x_{12} = b_1 \dot{-} b_1 = 0$ and $x_{21} = c_1 \dot{-} b_1$, as well as $x_{22} = e \dot{-} c_1 = c_2$ because $c_1 + c_2 = e$ and therefore $c_1 \leq e$. This shows that $x_{12} + x_{22} = c_2$ and $x_{21} + x_{22} = b_2$ by (8). Similarly, if $b_1 \not\leq c_1$, then $x_{21} = c_1 \dot{-} c_1 = 0$ and $x_{12} = b_1 \dot{-} c_1$, as well as $x_{22} = e \dot{-} b_1 = b_2$ because $b_1 + b_2 = e$ and therefore $b_1 \leq e$. This shows that $x_{21} + x_{22} = b_2$ and $x_{12} + x_{22} = c_2$ by (8).

8 Largest Consistency Witnesses

A key fact about standard relations is that if $R(X)$ and $S(Y)$ are two consistent standard relations, then there is a consistency witness $W(XY)$ for $R(X)$ and $S(Y)$ that is *largest* in the sense that every other consistency witness $U(XY)$ for $R(X)$ and $S(Y)$ is included in it, i.e., $U \subseteq W$ holds. This follows from the basic fact that if $W_1(XY)$ and $W_2(XY)$ are consistency witnesses for the standard relations $R(X)$ and $S(Y)$, then their set-theoretic union $W_1 \cup W_2$ is also a consistency witness for $R(X)$ and $S(Y)$. Therefore, the union of all consistency witnesses for $R(X)$ and $S(Y)$ is the largest consistency witness for them (and it actually coincides with the standard join $R \bowtie S$).

Assume that $\mathbb{K}$ is a positive commutative monoid and let $R(X)$ and $S(Y)$ be two consistent $\mathbb{K}$-relations. We say that a $\mathbb{K}$-relation $W(XY)$ is *a largest consistency witness for $R(X)$ and $S(Y)$* if for every consistency witness $U(XY)$ for $R(X)$ and $S(Y)$, we have $U' \subseteq W'$, where U', W' are the supports of $U(XY)$, $W(XY)$. In words, a largest consistency witness for two $\mathbb{K}$-relations is a consistency witness of largest support.

For arbitrary positive commutative monoids, largest consistency witnesses need not exist. A case in point is the bag monoid $\mathbb{N} = (N, +, 0)$. Specifically, consider the two bags $R(X) = \{a{:}1, b{:}1\}$ and $S(Y) = \{c{:}1, d{:}1\}$. These two bags are consistent, but their only two consistency witnesses are $W_1(XY) = \{ac{:}1, bd{:}1\}$ and $W_2(XY) = \{ad{:}1, bc{:}1\}$, which have incomparable supports. Consider also the positive commutative monoid $\mathbb{N}_2 = (\{0, 1, 2\}, \oplus, 0)$, where $1 \oplus 1 = 1 \oplus 2 = 2 \oplus 1 = 2 \oplus 2 = 2$, and 0 is the neutral element of $\oplus$. The same bags as above, but now viewed as $\mathbb{N}_2$-relations are an example of two consistent $\mathbb{N}_2$-relations with no largest consistency witness. Note that the monoid $\mathbb{N}_2$ is finite, while the monoid $\mathbb{N}$ is infinite.

Nonetheless, the property of standard consistent relations having largest consistency witnesses generalizes to relations over idempotent monoids, where a monoid $\mathbb{K} = (K, +, 0)$ is *idempotent* if the identity $x + x = x$ holds, for every $x \in K$.

PROPOSITION 3. *Let $\mathbb{K}$ be an idempotent and positive commutative monoid. Then, for every two consistent $\mathbb{K}$-relations, there is a largest consistency witness.*

Let $\mathbb{K}$ be such a monoid. If $R(X)$ and $S(Y)$ are two consistent $\mathbb{K}$-relations with consistency witnesses $W_1(XY)$ and $W_2(XY)$, then the $\mathbb{K}$-relation $T(XY)$ defined by $T(t) = W_1(t) + W_2(t)$ for every XY-tuple t is also a consistency witness for $R(X)$ and $S(Y)$. Indeed, for every X-tuple r and every Y-tuple s, we have

$$\begin{array}{l} T(r) = W_1(r) + W_2(r) = R(r) + R(r) = R(r) \\ T(s) = W_1(s) + W_2(s) = S(s) + S(s) = S(s). \end{array}$$

Therefore, $T[X] = R$ and $T[Y] = S$, so T is a consistency witness for R and S. We now claim that since $\mathbb{K}$ is positive and since $\mathbb{K}$-relations have (by definition) finite support, there is a consistency witness $W(XY)$ of largest support, which is then a largest consistency witness for $R(X)$ and $S(Y)$.

To see why this claim is true, suppose that $\mathbb{K}$ is positive and idempotent and that $R(X)$ and $S(Y)$ are consistent $\mathbb{K}$-relations. Let N be the number of tuples in the standard join of the standard relations $R' \bowtie S'$, where R' and S' are the supports of R and S. Since, by positivity, every consistency witness for R and S has its support in $R' \bowtie S'$, there are at most 2^N possible supports of witnesses of consistency for R and S. Let $M \leq 2^N$ be the number of such different supports and let $W_1, W_2, \ldots, W_M$ be a collection of witnesses of consistency such that their list of supports $W'_1, W'_2, \ldots, W'_M$ is the complete enumeration of all supports of witnesses of consistency. Now, consider the $\mathbb{K}$-relation $W(XY)$ defined by $W(t) = \sum_{i=1}^{M} W_i(t)$ for every XY-tuple t, where the sum is in $\mathbb{K}$. By the idempotency of $\mathbb{K}$, the $\mathbb{K}$-relation W is a consistency witness of R and S. And by the positivity of $\mathbb{K}$, the support W' of W contains the support W'_i of every W_i, and hence the support of any consistency witness for $R(X)$ and $S(Y)$ by the choice of the enumeration $W_1, W_2, \ldots, W_M$.

In addition to the Boolean monoid $\mathbb{B} = (\{0,1\}, \vee, 0)$, examples of idempotent monoids include the monoids $\mathbb{T} = (R \cup \{\infty\}, \min, \infty)$, $\mathbb{V} = ([0,1], \max, 0)$, and $\mathbb{P}(A) = (\mathcal{P}(A), \cup, \emptyset)$ introduced in Section 2.

Note that, unlike the case of standard relations, largest consistency witnesses need not be unique for relations over arbitrary idempotent monoids. To see this, consider the positive commutative monoid $\mathbb{L} = (Q^{\geq 0}, \max, 0)$ of non-negative rationals with maximum as operation, and 0 as neutral element, which is idempotent. Consider also the $\mathbb{L}$-relations $R(X) = \{a{:}1, b{:}1\}$ and $S(Y) = \{c{:}1, d{:}1\}$ with disjoint sets of attributes. These two $\mathbb{L}$-relations are consistent and have largest consistency witnesses, namely, any $\mathbb{L}$-relation $W(XY)$ of the form $\{ac{:}1, ad{:}p, bc{:}p, bd{:}1\}$ with $p \in (0,1]$ is a consistency witness for R and S. Thus, while the largest witnesses W are "canonical" in terms of *support*, they need not be "canonical" when taking the *annotations* into account.

There is another sense, however, in which idempotent positive commutative monoids admit canonical-looking consistency witnesses, in addition to having largest support. As discussed earlier, for every positive monoid $\mathbb{K} = (K, +, 0)$, there is a partial preorder $\leq$ on K defined by declaring that $x \leq y$ holds if and only if there exists $z \in K$ such that $x + z = y$. What holds for idempotent positive commutative monoids is that for every two consistent $\mathbb{K}$-relations $R(X)$ and $S(Y)$ and for every *finite* collection of consistency witnesses $U_1, \ldots, U_n$ there is a consistency witness $W(XY)$, still of largest support among all witnesses of consistency of R and S, such that $U_j(t) \leq W(t)$ holds in the preorder $\leq$ of $\mathbb{K}$, for every $j = 1, \ldots, n$ and every XY-tuple t. For this, simply ensure that all target witnesses $U_1, \ldots, U_n$ appear in the enumeration $W_1, W_2, \ldots, W_M$ featuring in the construction of W of the previous paragraph, perhaps by taking M to be an additive term n larger than it was, if necessary. Since by positivity the inequality $U_j(t) \leq \sum_{i=1}^{M} W_i(t) = W(t)$ holds whenever U_j appears in the enumeration $W_1, W_2, \ldots, W_M$, the claim follows. It is apparent from this argument that, in this construction, the witness W depends on the finite list $U_1, \ldots, U_n$; however, only the annotations depend on $U_1, \ldots, U_n$ and, therefore, W is still largest (with respect to support). As regards annotations, the set of consistency witnesses is *dense*: for every finite collection of consistency witnesses $U_1, \ldots, U_n$, there is a largest consistency witness W that sits simultaneously *above* all of them, point-wise in the preorder $\leq$, i.e., W satisfies $U_i(t) \leq W(t)$ for all $i = 1, \ldots, n$ and all XY-tuples t.

Furthermore, there is a case of special interest where not even the annotations of W need depend on the finite list $U_1, \ldots, U_n$. Specifically, if the monoid $\mathbb{K}$ is *finite*, then not only there is a finite number $M \leq 2^N$ of supports of consistency witnesses, but there is just a finite number of consistency witnesses overall. Thus, if in the construction of W we take M to be the total number of consistency witnesses and we let $W_1, W_2, \ldots, W_M$ to be the complete enumeration of these witnesses, then the resulting W is uniquely determined and sits simultaneously above *all* consistency witnesses.

Finally, we note that idempotency is not a necessary condition for the existence of largest consistency witnesses. Indeed, let $\mathbb{R}^{\geq 0} = (R^{\geq 0}, +, 0)$ be the monoid of the non-negative real numbers with addition. This monoid is the reduct of a semifield, namely, the semifield of non-negative reals with the standard addition and multiplication of real numbers, and the standard division by non-zero real numbers as inverse for the multiplication. The reduct $\mathbb{R}^{\geq 0}$ is a positive commutative monoid that has the transportation property but is not, of course, idempotent. Now consider two consistent $\mathbb{R}^{\geq 0}$-relations $R(X)$ and $S(Y)$. As with every other positive commutative monoid that arises from a semifield, their Vorob'ev join $W = R \bowtie_{\mathrm{V}} S$ as defined in Equation (6) is a consistency witness for R and S. It is easy to check that if $U(XY)$ is some other consistency witness, then $W(t) = 0$ implies $U(t) = 0$: indeed, by the absence of zero-divisors in any semifield, the multiplication in (6) gives $W(t) = 0$ only if $R(t[X]) = 0$ or $S(t[Y]) = 0$. Thus, by combining the positivity of the monoid with the fact that $U[X] = R$ and $U[Y] = S$, we get that $W(t) = 0$ only if $U(t) = 0$. This shows that $U' \subseteq W'$

and hence W has largest support. Indeed, in this case W is even "canonical" in its annotations because they depend only on R and S.

An open problem arising from the preceding discussion is to characterize the positive commutative monoids for which largest consistency witnesses always exist.

Acknowledgments Part of the research on this article was carried out at the Simons Institute for the Theory of Computing, where the authors were visiting in the spring and the fall of 2023. Atserias was partially supported by grant no. PID2022-138506NB-C22 (PROOFS BEYOND) and the Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in R&D (CEX2020-001084-M) of the AEI, and the CERCA and ICREA Academia Programmes of the Generalitat.

9 References

[1] A. Atserias and P. G. Kolaitis. Structure and complexity of bag consistency. In L. Libkin, R. Pichler, and P. Guagliardo, editors, *PODS'21: Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Virtual Event, China, June 20-25, 2021*, pages 247–259. ACM, 2021.

[2] A. Atserias and P. G. Kolaitis. Consistency of relations over monoids. *Journal of the ACM*, 72(3, Article 18):47 pages, 2025. Earlier version in Proc. ACM Manag. Data, 2(2):107, 2024.

[3] C. Beeri, R. Fagin, D. Maier, and M. Yannakakis. On the desirability of acyclic database schemes. *J. ACM*, 30(3):479–513, July 1983.

[4] K. M. Dannert, E. Grädel, M. Naaf, and V. Tannen. Semiring provenance for fixed-point logic. In C. Baier and J. Goubault-Larrecq, editors, *29th EACSL Annual Conference on Computer Science Logic, CSL 2021, January 25-28, 2021, Ljubljana, Slovenia (Virtual Conference)*, volume 183 of *LIPIcs*, pages 17:1–17:22. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[5] R. Fagin. Degrees of acyclicity for hypergraphs and relational database schemes. *J. ACM*, 30(3):514–550, 1983.

[6] E. Grädel and V. Tannen. Semiring provenance for first-order model checking. *CoRR*, abs/1712.01980, 2017.

[7] T. J. Green. Containment of conjunctive queries on annotated relations. *Theory Comput. Syst.*, 49(2):429–459, 2011.

[8] T. J. Green, G. Karvounarakis, and V. Tannen. Provenance semirings. In L. Libkin, editor, *Proceedings of the Twenty-Sixth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 11-13, 2007, Beijing, China*, pages 31–40. ACM, 2007.

[9] P. Honeyman, R. E. Ladner, and M. Yannakakis. Testing the universal instance assumption. *Inf. Process. Lett.*, 10(1):14–19, 1980.

[10] G. Karvounarakis and T. J. Green. Semiring-annotated data: queries and provenance? *SIGMOD Rec.*, 41(3):5–14, 2012.

[11] M. A. Khamis, H. Q. Ngo, R. Pichler, D. Suciu, and Y. R. Wang. Convergence of datalog over (pre-) semirings. *J. ACM*, 71(2):8:1–8:55, 2024.

[12] E. V. Kostylev, J. L. Reutter, and A. Z. Salamon. Classification of annotation semirings over containment of conjunctive queries. *ACM Trans. Database Syst.*, 39(1):1:1–1:39, 2014.

[13] R. Ramakrishnan. *Database Management Systems*. WCB/McGraw-Hill, 1998.

The Five Facets of Data Quality Assessment

Sedir Mohammed[1], Lisa Ehrlinger[1], Hazar Harmouch[2], Felix Naumann[1], Divesh Srivastava[3]
[1]Hasso Plattner Institute, University of Potsdam, Germany
[2]University of Amsterdam, The Netherlands
[3]AT&T Chief Data Office, USA
sedir.mohammed@hpi.de, lisa.ehrlinger@hpi.de
h.harmouch@uva.nl, felix.naumann@hpi.de, divesh@research.att.com

ABSTRACT

Data-oriented applications, their users, and even the law require data of high quality. Research has divided the rather vague notion of data quality into various dimensions, such as accuracy, consistency, and reputation. To achieve the goal of high data quality, many tools and techniques exist to clean and otherwise improve data. Yet, systematic research on actually assessing data quality in its dimensions is largely absent, and with it, the ability to gauge the success of any data cleaning effort.

We propose five facets as ingredients to assess data quality: *data*, *source*, *system*, *task*, and *human*. Tapping each facet for data quality assessment poses its own challenges. We show how overcoming these challenges helps data quality assessment for those data quality dimensions mentioned in Europe's AI Act. Our work concludes with a proposal for a comprehensive data quality assessment framework.

1 The Many Dimensions of Data Quality

Data quality (DQ) has been an important research topic for the last decades [10, 43, 62], reflecting its critical role in all fields where data are used to gain insights and make decisions. A manifold of DQ dimensions exists that regard data and their properties from various perspectives and contribute to understanding and characterizing the complex nature of data [10, 62].

The high demand for DQ. Especially in the fast-moving landscape of *artificial intelligence* (AI), where data plays a pivotal role, the significance of DQ is dramatically increasing, so much so that literature calls this trend a paradigm shift from a model-centric view to a data-centric one [64]. Data-centric AI emphasizes the data and their impact on the underlying model [44, 45, 63]. Literature showed that DQ, with its various dimensions, significantly influences prediction accuracy [24, 36, 40, 45]. Domain-specific particulars provide a context that imposes specific requirements on DQ assessment, such as

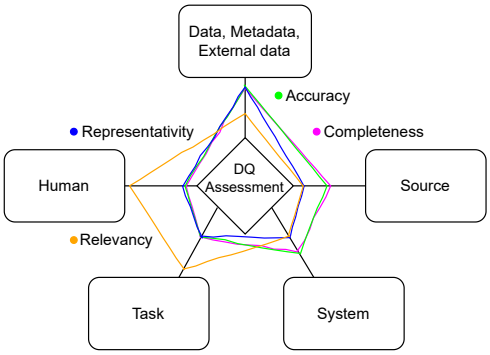


Figure 1: The five *facets* of DQ assessment and exemplary characteristics for DQ dimensions.

the *Health Insurance Portability and Accountability Act* (HIPAA), which focuses on privacy but promotes DQ dimensions, such as accuracy and completeness for ensuring trust [2].

Such requirements have also become part of regulation, as in the *General Data Protection Regulation* (GDPR) [25] and the EU *AI Act* [22]. For instance, the AI Act mentions in *Article 10* the DQ dimensions representativity, accuracy (free of errors), completeness and relevancy [22]. Similar initiatives to regulate DQ and AI are also being made by the United States [31] and China [52], which underlines the international interest in the topic of DQ.

Examining DQ is by no means just an academic problem [12]. Industry is also concerned about the impact of DQ on business [53]. Companies have shifted from internal "data gazing" [37] to hiring auditing firms for quality assurance. The literature shows that poor DQ has an enormous economic impact on organizations, either through loss of revenues or through additional internal costs [41, 50].

In addition to recognizing the relevance of DQ and understanding it in terms of the various dimensions, the goal is to improve DQ by cleaning the data. Yet, quality cannot be improved if it can-

not be measured [60]: we need concrete *assessment methods* to evaluate DQ in individual dimensions. Batini et al. [12] define DQ assessment as the measurement of DQ and the comparison with reference values for diagnosing it. As such, apart from the pure measurement of DQ, assessment includes classifying whether the measured quality is *sufficient* (or "fit") for the underlying task. *Measuring* vs. *judging* whether the measured DQ suffices for a task at hand are challenges of rather different natures.

Vision statement. Given a dataset, a use case (task specification), a set of DQ dimensions, and their formal definitions, our goal is to develop effective and efficient assessment procedures for each DQ dimension. These procedures should compute values that accurately align with the formal definitions.

Mission statement. To achieve the vision, we want to identify facets upon which assessment procedures across DQ dimensions depend. These facets enable individual dimensions to benefit from solutions to shared assessment challenges and streamlined implementation of assessment procedures.

Contribution. This paper proposes a new perspective on DQ research: through the lens of so-called *facets*. We discuss five *facets* of DQ assessment as potential sources for DQ information. Each *facet* presents its own set of challenges and opportunities. To overcome the challenges and capitalize on the opportunities, we identify a wide range of technologies that require cross-community expertise. We envision a thorough implementation of these technologies by different research communities. The ultimate goal is the integration of these technologies into a robust framework. We advocate for the development of a *DQ assessment framework* to accurately and efficiently measure all dimensions of the DQ. The framework enables (1) the integration of deeper data profiling methods [5], (2) compliance with given regulations, (3) enhancement of data cleaning, as well as (4) *judging* whether DQ meets user expectations. While this paper focuses on structured data, we believe it can also be extended to semi-structured or unstructured data.

2 Data Quality Assessment by Facets

Data quality assessment in its variety of dimensions [9, 43] poses many definitional, computational, and organizational challenges. We propose five *facets* (see Figure 1) that serve as foundation for DQ assessment: (i) the *data* itself, including metadata and external data; (ii) the *source* of the data; (iii) the *system* to store, handle, and access the data; (iv) the *task* to be performed on the data; and (v) the *humans* who interact with the data. These five facets are inspired by the stages of a typical data life cycle [59]: all relevant components of each stage can be mapped to one or more facets.

Each *facet* poses its own challenges and opportunities for future research. We hypothesize that addressing these challenges per *facet* adresses problems that arise from more than one DQ dimension. We propose *facets* as an additional layer to structure DQ research, allowing all dimensions involved in the assessment of a specific *facet* to benefit simultaneously from solving these challenges.

In the following, we define and discuss each of the five *facets* and their key challenges. We list exemplary DQ dimensions (see [39] for definitions) that specifically benefit from resolving these challenges.

2.1 The Data Facet

Raw data values are intended to represent real-world concepts and entities. The data facet includes the data semantics and their digital representation. It also includes metadata, such as schema information and other documentation, and any assessment-relevant external knowledge (as data), like a knowledge base (e.g., DBpedia [35]) to validate data. The data facet encompasses all challenges related to the data being assessed, its metadata, and external data.

As data occur in different *granularity* (e.g., values, records, columns), DQ assessment must identify the necessary level of detail and devise quality-metric aggregation methods to cross levels of granularity. Also, *metadata*, such as schema and data types, should be available and of high quality itself. When external knowledge is needed, challenges arise in discovering, matching, and assessing the quality of *reference data*. If data is encrypted, it cannot be assessed directly, so DQ assessment must handle *encrypted data* and, in case of distribution, also work in a *federated setup*.

In the following, we highlight two well-known DQ dimensions (mentioned in the AI Act) where the data facet is involved in the assessment.

Example DQ Dimensions

Accuracy: Typical metrics to assess accuracy require *reference data* to determine how closely the data matches the reality.

Completeness: Placeholders represent missing values, using either obvious placeholders like "NaN" or less obvious placeholders. The assessment needs *metadata* that contains information about the placeholder representation.

2.2 The Source Facet

The source of data represents a logical perspective. This *facet* encompasses evaluating the data generation and collection processes, as well as assessing the source's integrity and organizational compliance. The main aspect of the source facet is data *provenance*, which includes information on the origins, providers, and other organizations involved in creating and transforming the data [29].

One key challenge is ensuring *data lineage* traceability, including the data origin and its transformations [26]. Additionally, a *process-oriented view* is crucial, which includes evaluating the transformation process and the credibility of annotating agents in the DQ assessment. It is also important to consider the *time range* for assessing reliability over time; longer histories provide a more comprehensive view, while shorter intervals highlight recent changes.

Example DQ Dimensions

Reputation: The assessment requires evaluating a data source's credibility and reliability, and thus, considering historical reliability with *data lineage*.

Believability: The key challenge is to verify the data origin (*time range*), source transformations (*data lineage*), and involved entities (*process*).

2.3 The System Facet

The system facet pertains to a physical perspective, including the infrastructure and technology for storing, handling, and accessing the data. It also covers the system's technical compliance with legal and regulatory requirements, ensuring adherence to necessary data management standards.

The system facet raises challenges, such as *clarity* or *auditability*. The *clarity* includes documenting the system's architecture, data processing capabilities, interoperability with other systems, security features, and user interface aspects. *Auditability* is crucial to verify compliance with regulations, such as data deletion and security standards.

Example DQ Dimensions

Recoverability: Assessing the ability to restore a prior state of the data requires knowledge about the file system, backup procedures (*clarity*) and long-term storage regulations (*auditability*).

Portability: The key challenge is to understand the storage system, including file formats (*clarity*) and interoperability standards (*auditability*).

2.4 The Task Facet

The task facet pertains to the specific use case and the context in which the data is employed. Thus, it inherently aligns with the "fitness for use" definition of DQ [10, 62]. The task influences which parts of the data (e.g., columns, tuples) are considered and how well they represent the real world.

The task facet poses challenges regarding the *relevance* of the data, including the identification of relevant attributes and tuples. Also, the *risk* of the task, according to the AI Act, which defines minimal-, limited-, high- and unacceptable-risk AI systems, can determine the way DQ is assessed [1]. Higher risk categories require more stringent DQ assessment methods, including strict validation processes and documentation, to ensure compliance.

Example DQ Dimensions

Timeliness: The key challenge is defining an acceptable timeframe for tasks and to classify how long data are considered up-to-date or *relevant*.

Relevancy: The assessment involves balancing the need for complete information (*relevance*) against the risk of including unnecessary data that can violate legal requirements (*risk*).

2.5 The Human Facet

The human facet introduces a subjective view, while including the diverse groups that interact with the data, perform the task, and interpret the results. It aligns DQ with the specific needs and contexts in which users operate. Some DQ dimensions (e.g., relevancy, believability), require user surveys to assess experiences and challenges in handling the data. This subjective perspective makes it challenging to fully automate the assessment. The human facet presents challenges such as the need to *design surveys* that capture a range of expertise levels, or also the consideration of the *intent* of different user groups and their perspectives (e.g., developers, customers).

Example DQ Dimensions

Ease of manipulation: Since manipulability can impact accessibility positively and data integrity negatively, the assessment must consider the users *intent* of manipulation.

Relevancy: Determining relevant data varies by user perspective (*intent*). The evolving nature of relevancy with changing user needs, market trends, and legal standards complicates maintaining up-to-date assessments (*survey design*).

3 Facet Application

In the previous section, we listed example DQ dimensions per *facet*, for which the considered *facet* is involved in the assessment. Of course, the participation of the *facets* in assessing a DQ dimension occurs to varying degrees. We use a three-level system ("++", "+", "-") to indicate a *facets*' participation: "++" for strong involvement, "+" for medium, and "-" for low to no involvement. We determined the involvement of the *facets* through several discussion rounds among all authors until we reached a consensus. When determining the involvement of *facets*, we deliberately voted in favor of an objective and automatic assessment and thus tried to minimize the involvement of the human facet. Although DQ is often defined as "fitness for use" [62] the task facet is not necessarily included in the assessment.

In the following, we discuss the *facet* involvement and implications with respect to specific technologies for each DQ dimension from the AI Act: accuracy (free of errors), representativity, completeness, and relevancy [22] (see Figure 1). Additionally, we include a discussion on accuracy and relevancy as examples to illustrate why certain facets are not involved in the assessment.

3.1 DQ Dimension: Accuracy

Definition Accuracy describes the correspondence between a phenomenon in the world and its description as data [10].

Data	Source	System	Task	Human
++	+	+	-	-

The data facet is the primary contributor to the assessment of accuracy. Further aspects from the source facet (e.g., data provenance) and the system facet (e.g., storage technologies) are also relevant. Conversely, the task and human facets are less relevant: accuracy can be measured on a purely objective level, considering factual correctness and alignment with truth.

The literature established several metrics to assess accuracy [12, 28]. Most metrics require reference data, which corresponds to the data facet. To address this challenge, the reference data must be defined (e.g., its level of detail) and collected. Open data platforms, such as Kaggle [3] or general knowledge bases (e.g., Wikidata [4], DBpedia [35]), are well suited to collect a variety of data. To make use of such external data, they must be matched with the data using *schema matching* approaches [11, 19, 30, 49], which must handle different formats to process reference data from different sources [38]. This is particularly challenging with data that include *natural language*, demanding methods for semantic and syntactic processing, potentially using *large language models* [23].

In cases where access to such data platforms is too expensive or where no relevant data of sufficient quality could be found, *semantic web technologies* combined with *information retrieval approaches* would allow gathering data from the web, as external data for assessment [14, 27, 55].

In terms of the source facet, error detection and cleaning methods, such as NADEEF [18] or HoloClean [51], can be used to identify and correct data errors. The transformations applied must be clearly documented in the metadata (see Section 3.3).

The system in which the data is stored might be responsible for erroneous values caused by system failures, such as crashes or bugs. Thus, the system can lose information when saving new values, such as decimal points. Consequently, system robustness, data replication, and recovery processes must be included in the metadata. These aspects require a cataloging system to format the metadata in a machine-readable format (see also Section 3.3).

The system in which the data and metadata are located must ensure that access to them aligns with the relevant privacy provisions. If the data owner grants consent, where the consent information can also be part of the cataloging system, a partial decryption can be performed. Alternatively, encryption schemes such as *homomorphic encryption* can be used to assess and process the data/metadata while they are encrypted [6]. Compliance with privacy provisions is independent of the assessment of specific DQ dimensions.

3.2 DQ Dimension: Representativity

Definition Representativity aims to ensure that the characteristics of the reference data are present in the considered data [17, 33].

Data	Source	System	Task	Human
++	-	-	-	-

The data facet is the main contributor to the assessment of representativity.

Similar to accuracy, metrics to assess representativity require information on the reference data [15, 17]. Thus, the reference data must first be defined to establish a baseline for comparison in the assessment. In contrast to accuracy, assessing representativity does not require the complete reference data – summary statistics, respectively, data distributions of the attributes, are often sufficient. Depending on the data source, *metadata* may already contain

information about summary statistics and distributions. These metadata must be in a *structured format* (e.g., JSON or RDF) to enable automated access and further processing. Beyond uniform formatting, information must follow a *uniform schema* and *vocabulary* across data sources to ensure interoperability. The use of an *ontology* (e.g., Croissant [7] or DSD [21]) would ensure a standardized schema and vocabulary, improving interoperability.

Still, the data must be matched with the given data, even if it is in an aggregated format. But, *data matching* with less data is an easier task because there are fewer records and attributes to compare, reducing computational complexity and processing time. This simplifies schema matching, data cleaning, and handling diverse formats, leading to fewer errors and more straightforward and accurate matching criteria. Nevertheless, if the external data sources do not provide this information, the technologies the assessment requires to obtain and match the reference data overlap with the technologies mentioned in the context of accuracy.

3.3 DQ Dimension: Completeness

Definition Completeness refers to the extent to which data, including entities and attributes, are present according to the data schema [46].

Data	Source	System	Task	Human
++	+	+	-	-

When focusing on entry-level completeness, the data facet is primarily involved in the assessment; the source and system facets partially.

Since completeness represents the presence of the data, its assessment requires the measurement of missing values. While `null` or conventional placeholders like "NaN" for missing values are easily identified, more research is required to also identify so-called "hidden missing values" like "-99", "EMPTY", or default values [13, 48]. Identifying these hidden missing values can either be done through prior knowledge (in terms of metadata and sophisticated *Data Catalogs* [20] or, particularly suited for the ML context, with *Data Cards* [47]) or alternatively learned with ML models taking into account the context. Placeholders can differ for each data source or be domain-specific, which is why strict documentation is important. In addition, transformations on missing values, like deleted records or applied imputation strategies, must also be part of the metadata.

Similar to accuracy, the system in which the data is located might cause missing values, e.g., due to hardware failure. In the context of completeness, the system can lose data or fail to store new values, again necessitating metadata for recovery processes.

3.4 DQ Dimension: Relevancy

Definition Relevancy describes the extent to which the data are applicable and helpful for a given task [62].

Data	Source	System	Task	Human
+	-	-	++	++

While the task and the human facet mainly support the assessment of relevancy, the data facet is also involved. Conversely, the source and system facets are less relevant, as relevancy is solely determined by the data's usefulness for fulfilling a specific task, regardless of how or where it was created or stored.

To assess relevancy, stakeholders must *define* the given task, requiring domain experts to incorporate best practices and to understand the task's intricacies. Given the task, stakeholders and experts have to assess the relevancy of individual attributes and tuples. Alternatively, *statistical methods* can assess relevancy, e.g., Shapley or LIME calculate the feature importance to determine each feature's contribution to an ML model's prediction [56, 57, 61]. As feature importance is computationally complex, manual assessment might still be needed.

This manual assessment can be supported with *data profiling* [42] methods, comprising several tasks, such as, the automatic identification of distributions, functional dependencies, or data types. Based on the gathered information, experts can define domain- and task-specific criteria to assess the relevance of individual attributes and tuples using a *rating system* (e.g., Likert scale). Depending on the underlying task and its criticality, a larger-scale user study must be conducted to reflect various stakeholders and their perspectives. These surveys must follow the principles of good user *survey design* principles [34] and their creation should be independent from a given dataset to ensure an automated reuse for new or changed datasets.

4 Vision: A DQ Assessment Framework

In previous sections, we explored the challenges associated with different *facets* of DQ assessment and their applications to DQ dimensions. To promote this fresh look on DQ research, we envision a *DQ assessment framework* that implements the assessment methods along the *facets*. For instance, relevancy and timeliness intersect within the task facet: the specification of the downstream task (e.g., ML-based classification) determines whether the data

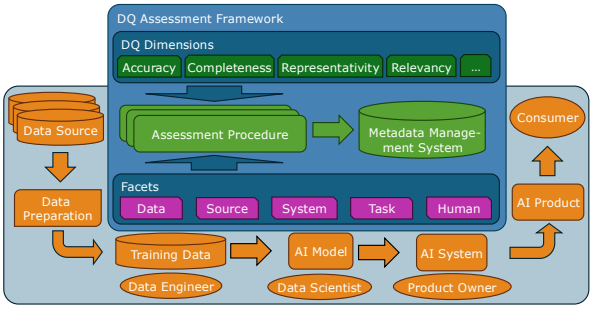


Figure 2: DQ assessment for an AI pipeline.

is relevant and also sufficiently up-to-date. The assessment of both dimensions benefits from that task specification.

Figure 2 shows the DQ assessment framework in the context of an AI pipeline. As part of this pipeline, data passes through various stages from its creation to the final product delivered to the customer. We can map the facets to these different stages of the pipeline. Thus, our proposed framework and the concept of facets are integrated into the AI pipeline: The data, in its digital representation (data facet) originate from various sources. A data engineer must prepare them using data preparation techniques, where all transformations must be traceable (source facet). The prepared data serve as training data, used by a data scientist to train an AI model, constituting a task (task facet). All these tasks can be deployed in an AI system (system facet), managed by a product owner, which in turn, can be part of an AI product that is delivered to customers. Finally, the various involved individuals should also be part of the DQ assessment (human facet). The assessment of each DQ dimension, together with the *facet's* participation, results in a dedicated assessment procedure.

We conducted an initial analysis of the participation of the *facets* per DQ dimension [39]. Apart from the facet-specific challenges to measure DQ in its various dimensions, building a framework that supports DQ measurement and management along the entire pipeline gives rise to further challenges:

Efficiency. The assessment effort and time should be low from a user perspective [8]. Data consumers might be unable or unwilling to wait for assessment results, and experts might not have much time to complete questionnaires or help in DQ assessment.

Explainability. Due to their ambiguity [32], assessment results must be explainable to consumers. In addition, the results should be traceable to their root cause, enabling measures to improve quality.

Metadata Management. Deploying the DQ assessment procedure requires an effective mechanism to store and query vast, diverse metadata (see *Metadata Management System* in Figure 2). An example solution and its challenges are discussed in [58].

5 Related Work

This section discusses representative works on DQ assessment and compares them to our fresh look through the lens of *facets*. Over the last decades, a number of DQ assessment frameworks have been proposed [12, 16]. For instance, Stvilia et al. [60] identified various sources for DQ assessment and distinguished intrinsic, relational, and reputational information quality. Batini et al. [12] divide the assessment into different phases and discuss metrics for DQ dimensions. Pipino et al. [46] present an approach combining subjective and objective DQ assessment results. In their vision paper, Sadiq et al. identify two dimensions to empirical DQ management [54]: the *metric* type (intrinsic vs. extrinsic) and the method *scope* (generic vs. tailored). They encourage the community to regard DQ beyond what we call the data facet – this paper follows that call. Other works [9, 10, 46] discuss challenges associated with specific DQ dimensions, e.g., the need for external data to assess accuracy [9].

In summary, many existing works implicitly mention individual facets (e.g., the human or the data facet) and the impact of their challenges on the assessment of DQ dimensions. However, so far, a unified view on how to address these different aspects was missing. We believe that addressing common DQ challenges per *facet* enables researchers the exploration of many DQ dimensions jointly.

6 Conclusion

We propose five assessment *facets* as foundational ingredients to assess *data quality* (DQ) and outline specific challenges and opportunities for each *facet*, highlighting the complexity of DQ assessment. We suggest how to overcome these challenges for the DQ dimensions mentioned in the AI Act as examples. Finally, we envision a DQ assessment framework that implements various methods to assess the DQ dimension through the lens of the *facets*.

Acknowledgements

This research was partially funded by the KITQAR project, supported by Denkfabrik Digitale Arbeitsgemeinschaft im Bundesministerium für Arbeit und Soziales (BMAS).

References

[1] EU AI Act: first regulation on artificial intelligence, 2023. URL https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. (Last accessed: 2024-07-25).

[2] HIPAA privacy rule to support reproductive health care privacy, 2024. URL https://www.federalregister.gov/documents/2024/04/26/2024-08503/hipaa-privacy-rule-to-support-reproductive-health-care-privacy. (Last accessed: 2024-07-25).

[3] Kaggle: Your machine learning and data science community, 2024. URL https://www.kaggle.com/. (Last accessed: 2024-07-15).

[4] Wikipedia, the free encyclopedia, 2024. URL https://www.wikipedia.org/. (Last accessed: 2024-07-15).

[5] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. Profiling relational data: a survey. *VLDB Journal*, 24(4):557–581, 2015. doi: 10.1007/S00778-015-0389-Y.

[6] Abbas Acar, Hidayet Aksu, A. Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys*, 51(4): 79:1–79:35, 2018. doi: 10.1145/3214303. URL https://doi.org/10.1145/3214303.

[7] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Pieter Gijsbers, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruyssen, Rajat Shinde, Elena Simperl, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Jos van der Velde, Steffen Vogler, and Carole-Jean Wu. Croissant: A metadata format for ml-ready datasets. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 1–6. ACM, 2024. doi: 10.1145/3650203.3663326. URL https://doi.org/10.1145/3650203.3663326.

[8] Donald P Ballou, InduShobha N Chengalur-Smith, and Richard Y Wang. Sample-based quality estimation of query results in relational database environments. *IEEE Transactions on Knowledge and Data Engineering*, 18(5):639–650, 2006.

[9] Carlo Batini and Monica Scannapieco. *Data quality: concepts, methodologies and techniques*. Data-centric systems and applications. Springer, 2006. ISBN 978-3-540-33172-8 978-3-642-06970-3.

[10] Carlo Batini and Monica Scannapieco. *Data and Information Quality: Dimensions, Principles and Techniques*. Springer Berlin Heidelberg, 2016. ISBN 978-3-319-24104-3.

[11] Carlo Batini, Maurizio Lenzerini, and Shamkant B. Navathe. A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys*, 18(4): 323–364, 1986. doi: 10.1145/27633.27634. URL https://doi.org/10.1145/27633.27634.

[12] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3):16:1–16:52, 2009. doi: 10.1145/1541880.1541883. URL https://doi.org/10.1145/1541880.1541883.

[13] Michal Bechny, Florian Sobieczky, Jürgen Zeindl, and Lisa Ehrlinger. Missing data patterns: From theory to an application in the steel industry. In *Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM)*, page 214–219, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384131. doi: 10.1145/3468791.3468841. URL https://doi.org/10.1145/3468791.3468841.

[14] Bettina Berendt, Andreas Hotho, and Gerd Stumme. Towards semantic web mining. In Ian Horrocks and James A. Hendler, editors, *Proceedings of the International Semantic Web Conference (ISWC)*, volume 2342 of *Lecture Notes in Computer Science*, pages 264–278. Springer, 2002. doi: 10.1007/3-540-48005-6_21. URL https://doi.org/10.1007/3-540-48005-6_21.

[15] Marcin Budka, Bogdan Gabrys, and Katarzyna Musial. On accuracy of PDF divergence estimators and their applicability to representative data sampling. *Entropy*, 13(7):1229–1266, 2011. doi: 10.3390/E13071229. URL https://doi.org/10.3390/e13071229.

[16] Corinna Cichy and Stefan Rass. An overview of data quality frameworks. *IEEE Access*, 7: 24634–24648, 2019.

[17] Line H. Clemmensen and Rune D. Kjærsgaard. Data representativity for machine learning and AI systems. *CoRR*, abs/2203.04706, 2022. doi: 10.48550/ARXIV.2203.04706. URL https://doi.org/10.48550/arXiv.2203.04706.

[18] Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed K. Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, and Nan Tang. NADEEF: a commodity data cleaning system. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 541–552. ACM, 2013. doi: 10.1145/2463676.2465327. URL https://doi.org/10.1145/2463676.2465327.

[19] AnHai Doan, Alon Halevy, and Zachary Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012. ISBN 978-0-12-416044-6. doi: 10.1016/C2011-0-06130-6. URL https://doi.org/10.1016/C2011-0-06130-6.

[20] Lisa Ehrlinger, Johannes Schrott, Martin Melichar, Nicolas Kirchmayr, and Wolfram Wöß. Data catalogs: A systematic literature review and guidelines to implementation. In *DEXA Workshops Proceedings*, volume 1479 of *Communications in Computer and Information Science*, pages 148–158. Springer, 2021. doi: 10.1007/978-3-030-87101-7_15. URL https://doi.org/10.1007/978-3-030-87101-7_15.

[21] Lisa Ehrlinger, Johannes Schrott, and Wolfram Wöß. Dsd: the data source description vocabulary. In *International Conference on Database and Expert Systems Applications (DEXA)*, pages 3–10. Springer, 2023.

[22] European Parliament. Artifical inteligence act. 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689. Version from 2024-06-13.

[23] Raul Castro Fernandez, Aaron J. Elmore, Michael J. Franklin, Sanjay Krishnan, and Chenhao Tan. How large language models will disrupt data management. *PVLDB*, 16(11):3302–3309, 2023. doi: 10.14778/3611479.3611527. URL https://www.vldb.org/pvldb/vol16/p3302-fernandez.pdf.

[24] Daniele Foroni, Matteo Lissandrini, and Yannis Velegrakis. Estimating the extent of the effects of data quality through observations. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1913–1918. IEEE, 2021. doi: 10.1109/ICDE51399.2021.00176. URL https://doi.org/10.1109/ICDE51399.2021.00176.

[25] GDPR. General data protection regulation (last accessed: 2024-02-13), 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504.

[26] Boris Glavic and Klaus R. Dittrich. Data provenance: A categorization of existing approaches. In *Proceedings of the Conference Datenbanksysteme in Business, Technologie und Web Technik (BTW)*, volume P-103 of *LNI*, pages 227–241. GI, 2007. URL https://dl.gi.de/handle/20.500.12116/31801.

[27] David A. Grossman and Ophir Frieder. *Information retrieval: algorithms and heuristics*. Number 15. Springer, 2nd ed edition, 2004. ISBN 978-1-4020-3004-8 978-1-4020-3003-1.

[28] Tom Haegemans, Monique Snoeck, and Wilfried Lemahieu. Towards a precise definition of data accuracy and a justification for its measure. In *Proceedings of the International Conference on Information Quality*, pages 16–16. MIT Information Quality (MITIQ) Program, 2016.

[29] Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. A survey on provenance: What for? what form? what from? *VLDB Journal*, 26(6):881–906, 2017. doi: 10.1007/S00778-017-0486-1. URL https://doi.org/10.1007/s00778-017-0486-1.

[30] Thomas N. Herzog, Fritz Scheuren, and William E. Winkler. *Data quality and record linkage techniques*. Springer, 2007. ISBN 978-0-387-69502-0. OCLC: ocn137313060.

[31] The White House. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, 2023. URL https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[32] Vimukthi Jayawardene, Shazia W. Sadiq, and Marta Indulska. The curse of dimensionality in data quality. In *Australasian Conference on Information Systems (ACIS)*, page 165, 2013. URL https://aisel.aisnet.org/acis2013/165.

[33] William Kruskal and Frederick Mosteller. Representative sampling, III: The current statistical literature. *International Statistical Review / Revue Internationale de Statistique*, 47(3): 245–265, 1979. doi: 10.2307/1402647.

[34] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research Methods in Human Computer Interaction*. Elsevier, second edition, 2017. ISBN 978-0-12-805390-4.

[35] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015. doi: 10.3233/SW-140134.

[36] Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, and Ce Zhang. Cleanml: A study for evaluating the impact of data cleaning on ML classification tasks. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 13–24. IEEE, 2021. doi: 10.1109/ICDE51399.2021.00009. URL https://doi.org/10.1109/ICDE51399.2021.00009.

[37] Arkady Maydanchik. *Data quality assessment*. Data quality for practitioners series. Technics Publications, 2007. ISBN 978-0-9771400-2-2.

[38] Tova Milo and Sagit Zohar. Using schema matching to simplify heterogeneous data translation. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 122–133, 1998. URL http://www.vldb.org/conf/1998/p122.pdf.

[39] Sedir Mohammed, Hazar Harmouch, Felix Naumann, and Divesh Srivastava. Data quality assessment: Challenges and opportunities. *CoRR*, abs/2403.00526, 2024. doi: 10.48550/ARXIV.2403.00526. URL https://doi.org/10.48550/arXiv.2403.00526.

[40] Sedir Mohammed, Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Sina Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. The effects of data quality on machine learning performance on tabular data. *Information Systems (IS)*, 132, 2025. doi: 10.1016/J.IS.2025.102549. URL https://doi.org/10.1016/j.is.2025.102549.

[41] Tadhg Nagle, Tom Redman, and David Sammon. Assessing data quality: A managerial call to action. *Business Horizons*, 63(3):325–337, 2020. ISSN 00076813. doi: 10.1016/j.bushor.2020.01.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0007681320300069.

[42] Felix Naumann. Data profiling revisited. *SIGMOD Rec.*, 42(4):40–49, 2013. doi: 10.1145/2590989.2590995. URL https://doi.org/10.1145/2590989.2590995.

[43] Felix Naumann and Claudia Rolker. Assessment methods for information quality criteria. In *Fifth Conference on Information Quality (IQ 2000)*, pages 148–162. MIT, 2000.

[44] Felix Neutatz, Binger Chen, Ziawasch Abedjan, and Eugene Wu. From cleaning before ML to cleaning for ML. *IEEE Data Engineering Bulletin*, 44(1):24–41, 2021. URL http://sites.computer.org/debull/A21mar/p24.pdf.

[45] Felix Neutatz, Binger Chen, Yazan Alkhatib, Jingwen Ye, and Ziawasch Abedjan. Data cleaning and automl: Would an optimizer choose to clean? *Datenbank-Spektrum*, 22 (2):121–130, 2022. doi: 10.1007/s13222-022-00413-2. URL https://doi.org/10.1007/s13222-022-00413-2.

[46] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002. ISSN 0001-0782. doi: 10.1145/505248.506010. URL https://doi.org/10.1145/505248.506010.

[47] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FaCCT)*, page 1776–1826, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3531146.3533231. URL https://doi.org/10.1145/3531146.3533231.

[48] Abdulhakim A. Qahtan, Ahmed Elmagarmid, Raul Castro Fernandez, Mourad Ouzzani, and Nan Tang. Fahes: A robust disguised missing values detector. In *Proceedings of the International Conference on Knowledge discovery and data mining (SIGKDD)*, page 2100–2109, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3220109. URL https://doi.org/10.1145/3219819.3220109.

[49] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001. doi: 10.1007/S007780100057. URL https://doi.org/10.1007/s007780100057.

[50] Thomas C Redman. *Data quality: the field guide*. Digital press, 2001.

[51] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. HoloClean: Holistic data repairs with probabilistic inference. *PVLDB*, 10(11):1190–1201, 2017. doi: 10.14778/3137628.3137631. URL http://www.vldb.org/pvldb/vol10/p1190-rekatsinas.pdf.

[52] Huw Roberts, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi. The chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI & SOCIETY*, 36(1):59–77, 2021. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-020-00992-2. URL https://link.springer.com/10.1007/s00146-020-00992-2.

[53] Shazia Sadiq, editor. *Handbook of data quality: research and practice*. Springer, 2013. ISBN 978-3-642-36256-9. doi: 10.1007/978-3-642-36257-6.

[54] Shazia Sadiq, Tamraparni Dasu, Xin Luna Dong, Juliana Freire, Ihab F. Ilyas, Sebastian Link, Miller J. Miller, Felix Naumann, Xiaofang Zhou, and Divesh Srivastava. Data quality: The role of empiricism. *SIGMOD Record*, 46(4):35–43, 2018. URL https://doi.org/10.1145/3186549.3186559.

[55] Urvi Shah, Timothy W. Finin, and Anupam Joshi. Information retrieval on the semantic web. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 461–468, 2002. doi: 10.1145/584792.584868. URL https://doi.org/10.1145/584792.584868.

[56] Lloyd S Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton, 1953.

[57] Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 9391–9404, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4e246a381baf2ce038b3b0f82c7d6fb4-Abstract.html.

[58] Divesh Srivastava and Yannis Velegrakis. Intensional associations between data and metadata. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 401–412. ACM, 2007. doi: 10.1145/1247480.1247526.

[59] Victoria Stodden. The data science life cycle: a disciplined approach to advancing data science as a science. *Communications of the ACM*, 63 (7):58–66, 2020. doi: 10.1145/3360646. URL https://doi.org/10.1145/3360646.

[60] Besiki Stvilia, Les Gasser, Michael B. Twidale, and Linda C. Smith. A framework for information quality assessment. *J. Assoc. Inf. Sci. Technol.*, 58(12):1720–1733, 2007. doi: 10.1002/ASI.20652. URL https://doi.org/10.1002/asi.20652.

[61] Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, pages 9269–9278. PMLR, 2020. URL http://proceedings.mlr.press/v119/sundararajan20b.html.

[62] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.*, 12(4):5–33, 1996. doi: 10.1080/07421222.1996.11518099. URL https://doi.org/10.1080/07421222.1996.11518099.

[63] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. Data collection and quality challenges in deep learning: a data-centric AI perspective. *VLDB Journal*, 32 (4):791–813, 2023. doi: 10.1007/S00778-022-00775-9. URL https://doi.org/10.1007/s00778-022-00775-9.

[64] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *CoRR*, abs/2303.10158, 2023. doi: 10.48550/ARXIV.2303.10158. URL https://doi.org/10.48550/arXiv.2303.10158.

Graph Data Management and Graph Machine Learning: Synergies and Opportunities

Arijit Khan[1] Xiangyu Ke[2] Yinghui Wu[3]
[1]Aalborg University, Denmark [2]Zhejiang University, China [3]Case Western Reserve University, USA
[1]arijitk@cs.aau.dk [2]xiangyu.ke@zju.edu.cn [3]yxw1650@case.edu

ABSTRACT

The ubiquity of machine learning, particularly deep learning, applied to graphs is evident in applications ranging from cheminformatics (drug discovery) and bioinformatics (protein interaction prediction) to knowledge graph-based query answering, fraud detection, and social network analysis. Concurrently, graph data management deals with the research and development of effective, efficient, scalable, robust, and user-friendly systems and algorithms for storing, processing, and analyzing vast quantities of heterogeneous and complex graph data. Our survey provides a comprehensive overview of the synergies between graph data management and graph machine learning, illustrating how they intertwine and mutually reinforce each other across the entire spectrum of the graph data science and machine learning pipeline. Specifically, the survey highlights two crucial aspects: **(1)** How graph data management enhances graph machine learning, including contributions such as improved graph neural network performance through graph data cleaning, scalable graph embedding, efficient graph-based vector data management, robust graph neural networks, user-friendly explainability methods; and **(2)** how graph machine learning, in turn, aids in graph data management, with a focus on applications like query answering over knowledge graphs and various data science tasks. We discuss pertinent open problems and delineate crucial research directions.

1. INTRODUCTION

Graph data, ranging from social and biological networks to financial transactions, knowledge bases, and transportation systems, permeates various domains. In these graphs, nodes represent entities with distinct features, while edges capture relationships between them. The growing volume of graph data and the increasing demand to extract value in real applications necessitate effective graph data management (GDM). Broadly speaking, data management encompasses a suite of algorithms and systems for acquiring, validating, storing, organizing, protecting, and processing data so they can be easily found and queried effectively, efficiently, securely, and cost-effectively. The principle of data management is to optimize data usage and comply with regulations, so to enable fair and responsible decision making, while maximizing the utility in downstream tasks. Modern data management challenges include the three V's of big data (volume, velocity, and veracity), dirty data, secure and distributed data processing, cloud computing, usability, new data types, emerging applications, etc. While general data management focuses on handling structured or semi-structured data such as tables and logs, graph data management presents unique challenges due to the interconnected nature of graph data. Managing relationships, traversals, and graph-specific queries (e.g., communities or reachabilities) demand specialized algorithms and data structures. Additionally, the irregularity and scale of graphs introduce challenges in indexing, storage, and real-time updates that go beyond traditional DM solutions. Specialized graph database management systems (graph DBMS), e.g., Neo4j, TigerGraph, Microsoft Cosmos DB, and Amazon Neptune were developed supporting graph transactions, queries, visualization, and diverse data models [113].

Machine learning (ML), a subfield of artificial intelligence (AI), uses algorithms to learn knowledge from data and generalize to unseen cases, often without explicit programming. Key principles of ML include data representation, performance evaluation on downstream tasks, and iterative optimization to improve accuracy. Based on the above requirements, ML models and systems are developed and deployed ensuring that they are effective, efficient, robust, and user-friendly. As ML becomes mainstream, there is a growing focus on explainability, transparency, fairness, safety, trust, and ethical decision-making. Graph machine learning (GML), in particular, graph neural networks (GNNs) have shown great promises for graph data-centric applications, such as classification, link prediction, community detection, question answering, and recommendation [135].

While data management (DM) and machine learning (ML) serve distinct purposes, their synergy is es-

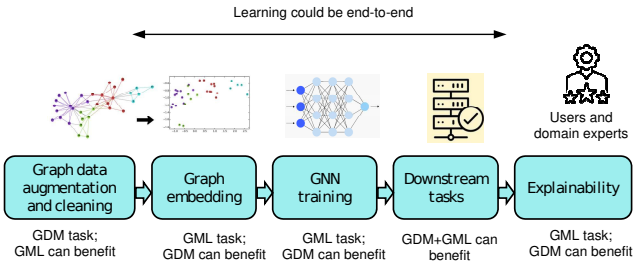


Figure 1: Graph data pipeline in data science and machine learning applications. Graph embedding can be task-specific or task-agnostic. Graph neural network (GNN) training can be end-to-end based on downstream tasks. We show which phases belong to GDM and which belong to GML, and can benefit from each other.

sential, as data is foundational to both. **First,** effective collaboration between data management and ML is necessary to unleash the full potential of an organization's data. For instance, data management techniques ensure clean, reliable, and up-to-date datasets, enabling ML models to generate accurate and trustworthy insights. **Second**, in modern data science applications, complex data undergo various processes involved in machine learning to generate the final predictive output, collectively forming a data pipeline [89, 59]. Figure 1 illustrates a representative graph data pipeline, encompassing the early stages of the graph data extraction, integration, cleaning, acquisition, validation, and enrichment; intermediate stages dealing with graph embedding, vector data, graph neural network (GNN) training, AutoML; and concluding stages involving downstream tasks and human-in-the-loop interactions, such as explaining the results of black-box GNN models. Managing effective and efficient data pipelines increases the need for robust data management solutions. **Third**, ML approaches enhance DM functionalities, e.g., ML can automate data transformation processes and might also understand a user's query intent to improve querying performance. Recent graph systems with ML capabilities [97, 39, 1] highlight the need to explore the synergies between two related fields: GDM and GML. Emerging technology landscapes such as AI, ML, edge computing, serverless and cloud computing, modern hardware, Internet-of-Things, data lakes, and Large Language Models (LLMs) are expanding the domain of data-driven downstream applications and what is feasible including real-time decision-making capabilities, streamlining integration, and enhanced security, making the synergy even more critical.

This survey examines the interplay between GDM and GML across different stages of the data pipeline depicted in Figure 1. We identify three key scenarios to structure the survey: **(a)** when GDM benefits GML; **(b)** when GML enhances GDM; and finally **(c)** when GDM + GML integration facilitates downstream tasks. For example, the initial phase of graph data cleaning is a GDM task, where we explore GML's contributions (§3.1). In contrast, stages like graph embedding, GNN training, and explainability focus on GML objectives, with GDM systems improving their efficiency and effectiveness (§3.2, §3.3, and §3.4). The fourth phase about downstream tasks benefits significantly from the synergy of GDM and GML, as discussed in §4.

Motivation: What are new in GDM and GML? With the rapid advances of graph machine learning (GML) techniques, such as graph embedding [137], GNNs [136, 155, 77], graph transformers [81], graphGPT [112], foundation models [72], and LLMs for graphs [47, 65], the role of graph data management (GDM) in the GML lifecycle has become increasingly vital. This spans all stages of the data pipeline, including preparation, improvement, embedding, training, and explanation. Recently, both academia and industry have emphasized the need for high-quality, large-scale data and robust, scalable, secure, and explainable models in ML systems [89, 167]. While there exist surveys and tutorials discussing the synergy between data management and ML – primarily focusing on relational data and relational database management systems [14, 59, 89], similar resources about graph data are comparatively scarce. Both GDM and GML pose significant challenges as follows.

In terms of GDM, graph data are inherently irregular, with nodes and edges forming complex, variable-length connections. This contrasts with the strict schema of relational data, where rows and columns provide a predictable structure. Therefore, specialized GDM systems are often required to quickly navigate and retrieve complex multi-hop neighbors. This places unique demands on GDM for efficient sampling and traversal strategies.

Due to their interconnected nature, partitioning graph nodes without disrupting critical structural properties, such as community boundaries, poses significant challenges. Unlike traditional data partitioning, where rows in tables can often be divided without impacting data relationships, graph partitioning must preserve inter-node dependencies to maintain the graph's integrity, typically requiring extensive communication between nodes or servers. GDM systems have to manage this inter-partition communication efficiently to support applications at scale – a challenge that is usually less pronounced in relational data management where tables can often be processed more independently.

Last but not least, graphs can be both high-dimensional and sparse, especially real-world graphs containing billions of nodes, but relatively few edges per node. Storing and efficiently retrieving meaningful patterns from these sparse yet high-dimensional graphs cause difficul-

ties that do not typically arise with dense tabular data.

Analogously, GML poses significant challenges due to the non-IID and unnormalized nature of graph data, the absence of strict schema, and irregular structures. Unlike traditional ML, where data samples are often processed independently, graph data require interdependent computations, leading to increased computational costs. Optimizing GML systems for model training, such as by supporting distributed training with efficient data loading and caching, is essential but challenging.

GML also generates high-dimensional embeddings for nodes and edges, which are crucial for tasks like node classification, link prediction, and similarity search. Managing these embeddings is more demanding than handling traditional ML data with simpler numeric or categorical features. Efficient storage, indexing, and retrieval mechanisms, such as vector databases or hybrid storage solutions, are essential for managing the large-volume high-dimensional embeddings produced by GML.

Additionally, reasoning in GML relies heavily on combining intricate feature interactions with graph topology. Graph data often require advanced feature engineering based on structural motifs or specific subgraph patterns. This demands robust support for pattern matching and subgraph extraction within GML systems, capabilities that are rarely needed in tabular ML. Scaling these operations for large graphs is particularly challenging and requires effective indexing and optimization strategies.

Finally, heterogeneity and multimodal graph data (e.g., graphs with nodes and edges having text and image-based features), along with emerging applications beyond classification and prediction (e.g., entity resolution [62], knowledge graphs question-answering [143], graph combinatorial optimizations [117]), and "black-box" deep learning approaches introduce further complexities to the deployment of GNNs.

Against this backdrop, our survey covering a set of the latest solutions that integrate GDM and GML techniques is both timely and relevant. We believe that our survey will attract and promote interdisciplinary research that advances scalable and explainable data pipelines for new data challenges in graph analysis.

Roadmap. In this survey, we demonstrate how graph data management and machine learning facilitate each other at different stages in a graph data pipeline. In particular, we delve into the following topics:

- Benefits of graph data cleaning and augmentation in improving the GNN performance (§3.1);
- Application of graph data management algorithms and systems for scalable graph embedding learning (§3.2);
- Vector data management using graph-based indexes (§3.3);
- GNN explainability methods, focusing on their usability and robustness (§3.4);
- Application of graph machine learning in knowledge graphs query answering (§4.1); and
- Applications of graph-based retrieval augmented generation (graph RAG) in large language models (LLMs) for data science tasks (§4.2).

We discuss background and related work in §2 and §5, respectively, and conclude with future work in §6.

2. BACKGROUND

We introduce background materials on graph neural networks and graph embeddings.

Graph neural networks (GNNs) are deep learning models to tackle graph-related tasks in an end-to-end manner [136]. GNNs have many variants, e.g., graph convolutional network (GCN) [57], graph attention network (GAT) [116], graph isomorphism network (GIN) [139], GraphSAGE [36], graph auto-encoder [56], graph generative adversarial network (GraphGAN) [119], and APPNP [58], etc. Specifically, graph convolution operations can be categorized as spectral [10] and spatial [21] approaches. In spectral methods, filters are applied on a graph's frequency modes computed via graph Fourier transform. Spectral formulations rely on the fixed spectrum of the graph Laplacian, and are suitable only for graphs with a single structure (and varying features on nodes), as well as are computationally expensive. On the other hand, spatial methods are not restricted to a fixed graph structure, as they extract local information by propagating features between neighboring nodes. Kipf and Welling [57] also develop a first-order approximation of the spectral convolution, which results in propagation between neighboring nodes. In particular, GCN adopts a general form as follows.

$$X^k = \delta(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}X^{k-1}\Theta^k) \quad (1)$$

Here $\hat{A} = A + I$, where I represents the identity matrix and A is the adjacency matrix of graph G. X^k indicates node feature representation in the k-th GCN layer, (with $X^0 = X$ a matrix of input node features). $\hat{D}$ represents the diagonal node degree matrix of $\hat{A}$, $\delta(.)$ is the non-linear activation function, and Θ^k represents the learnable weight matrix for the k-th layer. State-of-the-art GNNs follow a similar feature learning paradigm: Update the features of every node by aggregating the counterparts from its neighbors. The inference cost of feature propagation-based GNNs is usually polynomial-time [15, 58]. GNNs have been employed in node and graph classification (e.g., GCN [57], GAT [116], GraphSAGE [36], GIN [139]), link prediction (e.g., LGLP [12]), and entity resolution (e.g., GraphER [62]), etc.

Graph embedding or **representation learning** [11, 17] generates low-dimensional representation vectors of nodes, edges, and graphs that capture the structure and features of graphs accurately for downstream ML tasks.

Graph embedding algorithms can be categorized into three classes. **(1)** *Matrix factorization* methods [93] construct feature representations based on the adjacency or Laplacian matrix, and exploit spectral techniques. **(2)** *Random-walk* methods [32] transform a graph into a set of random walks via sampling and then employ Skip-Gram to generate embeddings. **(3)** *Graph neural networks* (GNNs)-based approaches [36, 116] focus on generalizing graph spectra into semi-supervised or supervised graph learning. They often follow a recursive neighborhood aggregation scheme to generate embeddings. State-of-the-art matrix factorization and random walk methods generally work on homogeneous graphs where nodes and edges share the same type, and the algorithms consider only graph structures. In contrast, GNN-based approaches exploit both graph structures and node features. They can be end-to-end, implying that the learning of embeddings is implicit within the GNN model and computed in a task-dependent manner. Embeddings of more complex networks such as heterogeneous information networks [107], relational graphs [102], hypergraphs [5], knowledge graphs [2], uncertain graphs [40], signed networks [150], dynamic graphs [7], spatio-temporal networks [105] have been studied.

3. GRAPH DATA MANAGEMENT FOR GRAPH ML

We discuss applications of graph data management such as data cleaning and augmentation in improving the GNN performance, graph algorithms, databases, and systems for scalable embedding learning, graph indexes for vector data management, and graph view-based explanation generation to enhance usability.

3.1 Graph Data Cleaning and Augmentation

Enhancing graph data to improve the performance of graph learning has seen an increased interest [167]. Existing data augmentation techniques from computer vision and natural language processing research cannot be easily generalized to irregular-shaped graph data. Graph data augmentation (GDA) [159] specifies enriching graph data to improve graph learning, which is categorized into "editing-based" and "representation-based". Editing-based methods aim to derive graph editing operations, such as removal, addition, or modify nodes, edges, features, or (sub)graphs [160, 162, 154, 37, 33], to improve the model performance such as graph neural networks. These methods may follow a deterministic process, learning to derive editing operations, or via a stochastic editing process. Graph sparsification [165], condensation [48], and diffusion [157] are also applied to improve GNN-based analysis.

Instead of deriving graph editing operators, representation based GDA directly learns to refine graph representation to improve follow-up analytical tasks. These methods train learnable parameters to generate augmented samples or graph representation and may adopt structure learning, adversarial training, contrastive learning, or automated augmentation [169, 161, 109]. Compared with representation-based approaches, editing-based GPA may be more interpretable and explainable, by performing data provenance analysis over the derived editing operators. On the other hand, representation-based approaches can be readily streamlined as input for downstream (graph) learning tasks, hence, may serve better in the need of end-to-end learning pipelines.

Error detection and repairing have been studied for graphs using rules and logic-based solutions, such as graph dependencies [26] and graph keys [24], neighborhood constraints [49, 68], uncertain edges cleaning [69]. Graph association rules (GARs) [25] detect missing links and semantic errors in graphs, while assisting in link prediction. GNNCleaner [138] repairs node labels to improve GNN robustness against label noise. Recent work such as SHACTOR [95] extracts validated shapes (a graph pattern carrying value, topological or cardinality constraints) with configurable measurements such as support to detect anomalies in knowledge graphs for error detection and cleaning. In general, rule-based error detection treats and deals with each error scenario in an isolated manner and often falls short of capturing complex scenarios where errors are from multiple sources with different forms, and may require additional effort to be adapted for general error detection.

Graph learning has been introduced to improve error detection and repairing for graphs. Generative adversarial learning and active learning has been exploited to improve error detection in graphs [33, 34]. For example, GALE [34] supports an interactive active, generative adversarial detection framework for graph error detection. The method applies few-shot learning to learn an error generation model that best fits a limited number of examples of different types of errors, and applies the model to augment the detected errors via a generative adversarial model to detect more errors. Active learning is adopted in this process to assist the error generation in the GAN-based error detection.

Synergy. There are good opportunities to integrate and interact with machine learning and graph data cleaning towards ML-based graph data cleaning systems. (1) Graph data constraints and rules can be exploited to characterize domain knowledge and context for ML data cleaning models. These graph data constraints and rules also provide a validation mechanism to make ML-based data cleaning reasonable. For example, graph association rules [25] or validation shapes [95] can be equipped with

learnable domain-specific patterns to improve the quality of domain-specific knowledge graphs. (2) The domain knowledge, context, and data constraints may also be properly featurized for potential training of foundational data cleaning models. The expressive ML models can be fine-tuned to perform downstream data cleaning tasks without conducting isolated, from-scratch data cleaning pipelines.

On the other hand, learning for graph error detection still requires a properly large amount of high-quality annotated examples, which remain a luxury for many applications such as domain sciences. Scaling ML solutions to large-scale graph cleaning also calls for efficient graph learning algorithms. Moreover, making ML-empowered data cleaning explainable with domain knowledge remains desirable yet a missing feature in current data systems. These provide opportunities for emerging needs such as fact checking tools in scientific knowledge graphs.

3.2 Scalable Graph Embedding and GNN Training

The surge of billion-scale graphs emphasizes the importance of efficient embedding learning on large graphs, as well as GNN training with them, such as for link prediction on Twitter with over one billion edges [35], users and products recommendation at Alibaba [120], etc. To scale GNNs to large graphs, various sampling strategies, e.g., node-wise sampling, layer-wise sampling, and graph-wise sampling are adopted [67].

To resolve efficiency and scalability issues with large graphs, recent works mainly focus on parallel computation, distributed systems, CPU-GPU hybrid architecture, and new hardware. PANE enables scalable and attributed networks embedding by measuring node attribute affinity with random walks, embedding computation via joint matrix factorization, and using multi-core parallelization [141]. DistGER exploits information oriented distributed random walks and distributed Skip-Gram learning for scalable graph embedding [27]. GraphVite [170] employs a CPU-GPU hybrid architecture, simultaneously performing graph random walks on CPUs and embedding training on GPUs. Marius [82] optimizes data movements between CPU and GPU on a single machine for large KG embedding. Seastar [134] develops a novel GNN training framework on GPUs with a vertex-centric programming paradigm. XGNN [111] designs a multi-GPU GNN training system to fully utilize GPU and CPU memory and high-speed interconnects. Amazon released DistDGL [166], a distributed graph embedding framework with mini-batch training using the Deep Graph Library (DGL). Facebook's Pytorch Biggraph [61] exploits graph partitioning and parameter servers to learn large-graph embeddings on multiple CPUs using PyTorch. ReGNN develops ReRAM-based architecture for GNN acceleration [71].

Synergy. Both graph embedding and GNN training are GML tasks. We showcase how GDM techniques can enhance them in four major ways: algorithms and systems, software-hardware co-design, and graph databases.

• *Efficient algorithms*. To improve efficiency and scalability of GNN training often at the cost of accuracy loss, mini-batch training and sampling strategies are developed, which can scale with data parallelism. Parallel and distributed training algorithms aim at reducing computation and communication overheads and design effective graph partitioning methods, all of which deal with irregularity, inter-connectedness, and sparseness in graph structure. Random walks approximate GNN message passing (e.g., APPNP [58]) and capture neighborhood structures for generating graph embedding. Therefore, improving the effectiveness of random walks, reducing their numbers and path lengths, as well as distributed random walk mechanisms have great potentials to improve the efficiency and scalability of GNN training and graph embedding. Efficient matrix factorization techniques can also gain superior performance and scale to embeddings of large-scale graphs.

• *Scalable systems*. Multi-CPU and multi-GPU platforms are widely-adopted scalable systems for distributed GNN training and graph embedding. Multi-CPU platforms enable distributed GNN training across multiple machines. Multi-GPU platforms employ CPU-GPU collaborative solutions, where GPUs conduct GNN training/ embedding, whereas CPUs handle computationally intensive tasks, including sampling, random walks, and workload partition. Modern hardware, e.g., FPGA, SSD, and ReRAM enable training larger graphs on a single machine, while providing accelerations, fault-awareness, and energy-efficiency.

• *Software-Hardware co-design*. PyTorch Geometric (PyG) and Deep Graph Library (DGL) are common software paradigms for GNN training. They support CPU and GPU computing, full-batch and mini-batch training, also provide APIs and user-defined functions to abstract computation and communication. Using them, more advanced software frameworks, e.g., AliGraph [168], DistGNN [115], and DistDGL [166] are developed which define user-friendly programming models (e.g., vertex-centric paradigm) and efficient data structures. They employ software-hardware co-design to reduce computation and communication costs via different parallelization schemes (e.g., pipeline parallelism), optimization strategies (e.g., synchronous vs. asynchronous communication, parameter server), on-chip data reuse, etc.

• *Graph databases*. Popular graph databases (graph DBs), e.g., Neo4J, ArangoDB, Amazon Neptune, TigerGraph, and Kùzu provide data science libraries and ML tools

to support a number of graph embedding methods and GNN training [52]. Graph DB's disk-based storage systems can be used with PyG remote backend to train a GNN model on very large graphs that do not fit on the main memory of a single server [1]. While graph DBs currently provide only basic graph ML functionalities such as node classification and regression, link prediction, it would be interesting to seamlessly integrate graph embeddings and GNN's capabilities into graph query processing and question answering (QA) (§4.1), also enabling vector indexes for efficient similarity search to facilitate graph RAG paradigm in LLMs (§4.2). These highlight the potential of graph DBs to be coupled with ML-based QA systems and LLMs [86, 73].

• *Improving graph data pipeline*. Finally, efficient graph embedding and GNN training are key to many downstream applications, e.g., graph data cleaning, entity resolution, and knowledge graph question answering, ensuring effective, efficient, and robust graph data pipelines.

3.3 Graph-based Vector Data Indexes

The management of vector data intersects with graph data management, particularly in systems that support graph-based machine learning (GML). A prime example is the use of graph-based vector indices, e.g., HNSW [20, 78, 28] to organize high-dimensional embeddings for retrieval tasks. These embeddings often originate from GML models like Graph Neural Networks (GNNs) [54, 9], where node or graph-level representations are computed for downstream applications. This synergy between graph-based indices and GML pipelines positions GDM systems, including Neo4j and TigerGraph, as comprehensive platforms for building GML workflows, integrating data storage, embedding generation, and similarity search functionalities.

Graph-based indices [84] diverge from traditional indexing methods, such as inverted indices [6, 46], locality-sensitive hashing [3, 114], and tree-based indices [8, 51], which typically partition vectors into buckets. Instead, graph-based indices construct proximity graphs, where nodes represent data points and edges denote neighbor relationships. These graph-based approaches present unparalleled effectiveness by leveraging semantic similarities through the principle that a neighbor's neighbor is likely to be a neighbor and iteratively expanding neighbors' neighbors through a best-first search [26, 124]. Recent works substantiate their scalability, positioning them for handling billion-scale datasets [127]. Unlike traditional graph data structures used for representing networked information, these indices are optimized for the Approximate Nearest Neighbor Search (ANNS) [4, 64, 127], a task foundational to many AI-driven applications. This makes them particularly relevant to GDM systems that serve as infrastructure for hybrid tasks combining traditional graph analysis and ML-based embedding retrieval. The implications of such methods extend beyond ANNS, permeating into the fabric of LLMs [54] and unstructured data management [41, 131], heralding a new era in the intersection of graph-based data management and real-world applications.

Graph-based vector indices have been subject to a range of optimizations aimed at improving both the index structure and search procedures, which can be categorized into four key areas:

The first major category, *graph index optimization*, focuses on diversifying neighbor connections to enhance graph navigability and capture semantic relationships between embeddings, such as refining the quality of the edge set [28], leveraging more sophisticated distance functions [20], adaptive neighbor selection [88], and hierarchical layouts [78]. These ensure that similar embeddings are efficiently connected and easily discoverable during search. The index graph quality directly impacts downstream GML tasks like node classification and link prediction, where effective and efficient similarity assessments are critical for model performance.

The second set of optimizations focuses on enhancing search strategies to reduce traversal overhead while maintaining high query accuracy. This is particularly important when scaling graph-based models to larger datasets, as the cost of inefficient traversal can quickly overwhelm the benefits of an optimized index. *Routing optimizations* address this challenge by refining key aspects such as entry point acquisition [74, 164], routing strategy [30, 151, 75, 151], and termination conditions [63, 152]. By combining these strategies, routing optimization ensures that even in large-scale GML datasets, searches remain fast and precise, minimizing the impact of increasing data size on performance.

Building on these search optimizations, the third category focuses on scaling solutions through *hardware-aware optimizations*, which adapt the index layout and search strategies to specific hardware capabilities [126]. Graph-based methods have been implemented in *external memory* such as heterogeneous memory (HM) [99] and solid-state disk (SSD) [45], to scale the system beyond traditional memory limitations. A recent work [44] has adapted graph-based indexes to the cutting-edge compute express link (CXL) architecture. In addition, *acceleration hardware* such as GPUs and FPGAs are utilized to parallelize vector computation [163, 83, 79] or data structure maintenance [146], providing an order of magnitude increase in efficiency for both index construction and search, These innovations exemplify how *software-hardware collaboration* enables scalable solutions for embedding-intensive GML tasks, addressing computational bottlenecks in GML workflows.

[1] https://blog.kuzudb.com/post/kuzu-pyg-remote-backend/

The fourth line of research integrates additional information into graph indices to further support more sophisticated retrieval scenarios, a critical need for complex graph ML workflows. Techniques such as attribute-based filtering [123, 31, 87, 171] incorporate structured attributes directly into the index, enabling hybrid queries that combine structured and unstructured data. For instance, in *multimodal search* scenarios, where each entity comprises multiple vectors, *multiple* graph indexes may be constructed and scanned to address a multi-vector query [121, 152]. An innovative approach [122] has fused multiple embeddings into a unified graph index with automatic weight learning, enabling efficient and accurate multimodal queries. These methods have demonstrated applicability in ML-powered systems, such as LLM-based online query answering [125], further bridging the gap between advanced data management and real-world applications.

Synergy. We illustrate how advancements in graph-based vector indices, a core GDM technique, significantly contribute to the scalability and efficacy of GML systems.
• *Efficient embedding management.* Graph indices excel at managing high-dimensional embeddings generated by GML tasks, such as node classification and link prediction. By leveraging optimizations in graph structure and search procedures, including neighbor diversification, efficient routing, and hardware acceleration, these indices enable faster and more precise similarity searches essential for embedding-driven GML workflows.
• *Scalable multimodal integration.* For multimodal GML tasks, where nodes or entities are represented by multiple embeddings, graph indices adapt to efficiently handle hybrid and multimodal queries. Techniques like fused graph indices allow simultaneous processing of multiple data modalities, directly benefiting use cases like multimodal knowledge retrieval and enhanced representation learning in large-scale systems.
• *Hardware acceleration for GML.* The alignment of graph indices with emerging hardware architectures, such as GPUs, FPGAs, and CXL, drives substantial improvements in computation and memory efficiency. These optimizations enable graph ML systems to scale effectively, overcoming the limitations of traditional memory-based approaches for embedding-intensive workloads.
• *Enhanced machine learning pipelines.* By integrating attribute filtering, handling incomplete data, and accommodating large-scale retrieval, graph indices bolster the robustness of GML pipelines. This ensures reliable and efficient data processing for tasks such as hybrid query answering, anomaly detection, and fair representation learning. The adaptability of graph indices to evolving GML requirements demonstrates their critical role in enabling complex, real-world applications.

3.4 GNN Explainability

To safely and trustfully deploy deep neural models, it is critical to provide human-intelligible explanations to end users and domain experts: *Which aspects of the input data drive the decisions of the model?* Therefore, explainability methods for GNNs are becoming popular.

Deriving and comparing GNN explanations are difficult. **(1)** There is no unique notion of explainability – the requirements arise due to many factors, e.g., trust, causality, transferability, fair decision making, model debugging, informativeness, etc. [70, 55]. **(2)** Analogously, several quantitative metrics such as fidelity, sparsity, contrastivity, and stability are proposed to evaluate explanation quality. It may be required to modify these metrics to capture the complex dependency of structure and feature in the graph space. For instance, perturbation-based metrics (e.g., fidelity) can drastically change the graph's structure, resulting in data outside the training distribution. Instead, a standard practice is to consider "milder" perturbations by removing associated features of important nodes and edges, while keeping the graph structure intact [90, 148]. **(3)** Due to the emerging nature of graph data and downstream tasks, there has been less qualitative evaluation of GNN explainability (e.g., human grounded evaluation) [118, 96]. Lack of real-world ground-truth explanations, complexity in graph data, and requirement of expert domain knowledge are key bottlenecks behind qualitative evaluation. **(4)** The output of GNN explainability (e.g., nodes, edges, features, subgraphs) and their categories (e.g., factual vs. counterfactual, instance vs. model-level) are different. For a holistic evaluation, such factors must be considered [53]. **(5)** Other concerns include non-robust GNN models and training bias [23].

Recently, many explainability methods for GNNs have been developed, which can be categorized across several dimensions [148, 50, 53]. *Self-explanatory* approaches incorporate explainability directly into GNN models, e.g., [18, 156]. *Post-hoc* methods [144, 29, 76, 147, 101, 118, 149] create a separate model to provide explanations for an existing GNN. In *global* explanation methods, users understand how the model works globally by inspecting the structures and parameters of a GNN model, or by generating graph patterns which maximize a certain prediction of the model [147]. In contrast, *local* methods examine an individual prediction of a model, figuring out why the model makes the decision on a specific test instance [144, 29, 101, 118, 149]. *Forward* explainability methods are GNN model-agnostic by learning evidences about graphs or nodes passed through the GNN. They can be *perturbation-based*, that is, masking some node features and/or edge features and analyzing the changes when the modified graphs are passed through GNNs [144]. They might also employ a simple, explain-

able *surrogate model* to approximate the predictions of a complex GNN [118]. In contrast, *backward* interpretability methods are GNN model-specific and can be either *gradient-based* [90] – backpropagating importance signals backward from the output neuron of the model to the individual nodes of the input graph, or *decomposition-based* [103] – distributing the prediction score in a backpropagation manner until the input layer. Thus, one identifies which nodes, edges, and features contribute the most to the specific output label in the GNN. Furthermore, GNN explainability methods can be classified as *factual* (i.e., finding a subgraph whose information is sufficient – which, if retained, will result in the same prediction), *counterfactual* (i.e., finding a subgraph that is necessary – which, if removed, will result in a different prediction), or both [110].

However, existing approaches in this field are limited to providing explanations for individual instances or specific class labels. The main focus of these methods is on defining explanations as crucial input features, often in the shape of numerical encoding. These methods generally fall short in *providing targeted and configurable explanations for multiple class labels of interest*. Additionally, existing methods may return large explanation structures and hence are not easily comprehensible. These explanation structures often lack direct accessibility and cannot be queried easily, posing a challenge for expert users who seek to inspect the specific reasoning behind a GNN's decision based on domain knowledge.

A recent work, GVEX [16] proposes a novel two-tier explanation structure called *explanation views*. An explanation view (similar to *graph view*) comprises a collection of graph patterns along with a set of induced explanation subgraphs. Given a database of multiple graphs and a specific class label assigned by a GNN-based classifier, lower-tier subgraphs provide insights into the reasons behind the assignment of the label by the classifier. They serve as both factual (that preserves the result of classification) and counterfactual explanations (which flips the result if removed). On the other hand, the higher-tier patterns summarize the subgraphs using common substructures for efficient search and exploration of these subgraphs. Analogously, RoboGExp [92] introduces a new class of explanation structures to provide robust, both counterfactual and factual explanations for graph neural networks. Given a GNN, a robust explanation refers to the fraction of a graph that are counterfactual and factual explanation of the results of the GNN over the graph, but also remains so for any "disturbance" by flipping up to k of its node pairs. In particular, such explanation indicates "invariant" representative structures for similar graphs that fall into the same group, i.e., be "robust" to small changes of the graphs, and be both "factual" and "counterfactual". Both GVEX and RoboGExp also emphasize effective, efficient, and scalable explanation generation by providing theoretical approximation guarantees and developing parallel and streaming algorithms.

Synergy. We depict how GDM assists in generating better GNN explanations, which is a GML task.

• *Useful explanations*. First, explanations should not only dissect the decision-making process of GNN models, but can also *zoom in/out* on how certain features, nodes, or subgraphs contribute to specific classifications, that is, explanations can be provided across multiple granularity of concept hierarchy depending on the needs of end users. Moreover, enhancing the *accessibility*, *configurability*, and *queryability* of explanations is crucial. Graph view-based two-tier explanations in GVEX [16] provide the first step in this direction, and a natural extension might be generating an explanation OLAP cube that can be drill up/down based on domain-specific requirements. Second, explanations should be presented in a *user-friendly* manner, possibly through visualizations or interactive tools that allow users to explore and interrogate GNNs' decisions. These tools could enable desirable capabilities, e.g., highlighting critical substructures, providing interactive interfaces, and allowing tunable parameters for domain experts to "query" the model about its decisions. It is paramount to think beyond "explanation of GNN models" and towards "explanations for users" to enable trust and effective deployment.

• *Efficient explanations*. Past research on explanation generation often does not emphasize on efficiency and scalability, e.g., requiring more than one day to generate an explanation over large-scale graphs [16]. Real-time explanations are key to interactiveness, configurability, queryability, and in-depth exploration of GNNs' decision making process. Parallel, streaming, and anytime algorithms, modern hardware, and software-hardware co-design have potentials to reduce explanation time.

• *Diversified explanations*. As stated earlier, several quantitative metrics, e.g., fidelity, sparsity, contrastivity, and stability are designed to evaluate explanation quality; however, no single measure is the best. It is important to pursue explanations that optimize multi-objective quality criteria, while also improving diversity. Concepts from databases, such as Pareto optimality and a skyline set of explanatory subgraphs can be useful.

• *Better explanations to improve data pipeline*. Finally, explanations can reveal unfairness in GNN's decision making process, detect anomalies and potential threats, help in model debugging, and assist organizations in meeting compliance and regulations, thereby improving the robustness of graph data pipeline.

4. GRAPH ML FOR GRAPH DATA MANAGEMENT

We illustrate applications of graph machine learning and graph-based LLMs in knowledge graphs query answering and other data science tasks.

4.1 Knowledge Graphs Query Answering

Query answering over datasets is an important data management task. We consider knowledge graph (KG) – a graph-based data model to store facts – denoted as ⟨subject, predicate, object⟩ triples, or a large-scale graph having nodes (subjects and objects) and edges (predicates) [132]. Querying KGs is critical for web search, semantic search, fact checking, and personal assistants. However, it is difficult due to their massive volume, heterogeneity, incompleteness, and schema flexibility. Additionally, a user's query (e.g., natural language query or query graph) may not match exactly w.r.t. entities, relations, and structure of the KG, requiring approximate matches to retrieve relevant answers [52].

Machine learning assists in **(1)** inferencing over KGs to identify missing relations during query answering, and also **(2)** finding approximate matches for queries [1, 39, 97]. **(3)** Natural language queries (NLQs) are semantically parsed to structured queries (e.g., SPARQL queries over KGs) using neural approaches [94]. **(4)** More recent techniques employ sequential models for end-to-end answering of NLQs over KGs, e.g., KEQA [42] for simple NLQs and EmbedKGQA [100] for multi-hop NLQs. **(5)** KG embedding methods can be useful as well. Wang et al. [129, 128] decompose multi-hop and complex queries into smaller subqueries, answer each subquery via single-hop reasoning with KG embedding, and then assemble the answers. In contrast, Query2box [98] and follow-up works train on multi-hop queries – they embed multi-hop logic queries and their answers (i.e., entities from a KG) in the same embedding space to reduce the query processing cost via inference.

Domain-specific knowledge graphs (KGs) [124, 66] have been curated to host scientific, factual knowledge rather than generic Web or common knowledge, such as KGs in material science, healthcare, medicine, education, cybersecurity, biology, and chemistry. While knowledge curation has been extensively studied, searching domain data remains nontrivial. Domain experts are still expected to write complex declarative queries (such as SPARQL), or data scripts to access KGs. There is a gap between the need of accessing KGs with (domain) languages and optimized query processing within state-of-the-art KG data systems. The rise of large language models (LLMs), such as GPT provides promising capabilities in generating natural language solutions in response to users' prompts. There are efforts on linking LLMs to KG search and exploration [85], as well as LLM-based knowledge graph exploratory search [60]. KG-enhanced, LLM-based QA is also studied: QAGNN [143] and GreaseLM [153] fine-tune a vanilla LM with a KG on downstream tasks, whereas DRAGON [142] and JAKET [145] perform self-supervised pre-training from both text and KGs at scale.

Synergy. Query processing is the bread-and-butter for the data management community. We highlight how GML and LLMs assist in KG querying and QA.

• *Natural language query processing*. Natural language interfaces to databases (NLIDB) is the holy grail for query interface to DBs – automatically translating natural language questions (NLQs) to structured queries (e.g., SQL) that can be processed by a database management system. With the prevalence of graph data (e.g., domain-specific KGs) and the standardization of graph query languages (GQL), there is an emerging need to covert NLQs to graph queries, e.g., Cypher, SPARQL, Gremlin, GSQL, PGQL, etc. This is more challenging due to the complexity and expressivity of graph queries, coupled with the schema-flexibility and heterogeneity in graph data. GNNs and LLMs can assist in these tasks because of their understanding of contexts in conversational QA, background knowledge, and capability of dealing with natural language text. For instance, Neo4J recently developed NeoDash[2] which leverages LLMs to interpret user's input NLQs and generates Cypher queries based on the provided schema definition.

• *Approximate query processing*. KGs are schema flexible, i.e., similar relationships between entity pairs can be represented in different ways. Therefore, one needs to construct various query patterns to retrieve all relevant answers from the underlying dataset, which is challenging. This necessitates approximate matches w.r.t. users' queries by understanding the query intent – KG embedding and KG + query embedding approaches can support approximate matching via inference.

• *Query processing over incomplete data*. KGs follow the open-world assumption, i.e., they are incomplete. To retrieve the complete set of answers for a given query, one must infer missing relations in KGs. In contrast, relational DBs generally follow the closed-world assumption with the presumption that all relevant knowledge is explicitly stored within the DB. Additionally, dealing with missing graph structure is more challenging than imputing missing feature values. ML-based link prediction and multi-hop inference techniques can be coupled with graph queries to resolve these problems.

• *Multimodal and multilingual data and queries*. Entities and relations in a KG can have features with different data modalities, e.g., text, images, and multimedia data. Analogously, text data in node features and

[2] https://neo4j.com/labs/neodash/2.4/user-guide/extensions/natural-language-queries/

queries can be in different languages. Dense vector embedding of multimodal and multilingual data, obtained via deep models, provide a unique opportunity to query such heterogeneous data. Data management techniques can also contribute in querying vector data with high-dimensional indexes and join, leveraging modern hardware and geometric data processing.

• *Graph databases and query optimization*. ML approaches, e.g., deep learning, reinforcement learning, and LLMs have shown promises in optimizing database queries and enhancing database administration functions such as query optimization, workload management, indexing, and storage layouts. Although there are recent developments in deep learning methods for graph pattern search and cardinality estimation [158], more work is needed in AI-facilitated graph databases and query optimization. Graph ML algorithms could play a pivotal role in predicting access patterns, node importance, learning graph indexes based on query characteristics. By leveraging historical usage data and graph topology, these systems can autonomously adapt storage strategies and retrieval mechanisms to match the evolving needs.

4.2 Graph RAG-based LLMs in Data Science Applications

LLMs which are a category of generative AI models and proficient at generating new text contents, offer a myriad of opportunities in data science by automating data analysis, manipulation, querying, and interpretation, as well as in code synthesis, digital assistants, finance, law, and education. Nevertheless, due to poor reasoning capacity, outdated or lack of domain knowledge, expensive re-training costs, and limited context lengths of LLMs, LLM-based data science pipelines often struggle with complex tasks – they hallucinate, i.e., generate factually incorrect, or even harmful contents. To address these issues, KGs are used as background knowledge to enhance LLMs for downstream tasks. The questions are parsed to identify relevant subgraphs from KGs, then they are integrated and fused with LLMs based on knowledge integration, prompt augmentation, and retrieval augmented generation (RAG). This framework, known as graph RAG or KG-RAG [140], is increasingly becoming popular due to its ability to capture the global context, compared to conventional RAG that retrieves knowledge from embeddings of textual chunks.

Recent works [80, 38, 130, 133] develop KG-unified language models in a graph RAG style. They can be broadly categorized into two groups according to the roles of KGs: **(1)** KGs as background knowledge, and **(2)** KGs as reasoning guidelines. While the former only retrieves relevant subgraphs as contexts based on input questions, the later retrieves the most relevant paths adaptively to guide the LLM's reasoning process [106]. Graph RAG is further added within LLM-based agent systems to leverage structured knowledge for enhanced decision-making and problem-solving capabilities [108].

Synergy. Besides GDM, effective text or vector processing may benefit graph RAG. For example, **(1)** What is the proper data model to represent and feed the retrieved knowledge to the LLM? Options include prompt-based or embedding-based data model. For the former, prompt engineering can be explored, such as serializing subgraphs to token sequences or ⟨subject, predicate, object⟩ triples, to best exploit LLMs' ability of text (natural language) processing. The latter can be better supported by vector databases (see § 3.3). **(2)** How to design indexes, search algorithms, and systems for more complex and hybrid vector search, including graph traversal with vector retrieval? Those may require unifying graph DBs and vector DBs as external memory of LLMs. **(3)** Graph query optimization, (explanatory) views, and provenance can help in making graph RAG better grounded by linking LLM response to factual knowledge at scale. **(4)** Last but not least, graph DBs may be used as "semantic caches" of LLMs by indexing previous question-answer pairs into a graph or vector space, enabling semantic matching with new queries instead of more expensive LLM API calls. These create new opportunities for GDM and broader data management techniques to play critical roles for graph RAG systems.

5. RELATED WORK

The closest to our work are surveys and tutorials on ML for data management and data management for ML, emphasizing on relational data and RDBMS [14, 59, 89, 43]. However, graph data result in unique challenges to both data management and ML (§1), justifying the importance of our survey.

Additionally, there are related surveys and tutorials on, e.g., graph representation learning [11, 17], graph neural networks [136, 155, 77], AI for data preparation [13], the role of graph data in graph ML [167], distributed GNN training [104], explainable AI in data management [91], ML explainability and robustness [19], LLM+KG [85], and high-dimensional vector similarity search [22], etc. However, none of them investigate the synergy of GDM and GML. To the best of our knowledge, ours is the first survey exploring the synergies between graph data management and graph ML over the end-to-end graph data pipeline. We hope that our survey will bridge the gap between these two popular domains – GDM and GML, and would inspire others to work on the emerging graph data challenges at their intersection.

6. FUTURE DIRECTIONS

Future work can be in several directions.

Real-time Graph Learning and Inference. The integration of spatiotemporal GNNs and dynamic graphs would enable the real-time decision that can rapidly explore evolving nodes and links. This calls for adaptive graph query processing and optimization, online graph learning, and real-time inference at scale. Graph analysis in finance, healthcare, security, and manufacturing will benefit significantly from this capability.

Privacy-preserving Graph ML. As the usage of graph data expands, so does the concern for privacy and security. Future developments in the synergy between graph machine learning and data management could delve into advanced privacy-preserving techniques for graph data. This might involve the integration of federated learning approaches, differential privacy, or novel encryption methods tailored to the unique characteristics of graph structures. Ensuring the confidentiality of sensitive graph information, while still extracting valuable insights, poses an exciting challenge.

Robust Graph ML. GNNs can be sensitive under a set of link perturbations or adversarial attacks. ML communities have investigated several approaches on how to quantify and improve the robustness of graph learning, e.g., certifiable robustness. Data management techniques such as graph sparsification and cleaning can also be employed. In the past, data imputation and integration for graphs have been extensively studied with the objective of data correctness and completeness, instead it would be interesting to clean graphs for optimizing the robustness of graph learning.

Unifying LLMs+KGs+Vector DBs. Knowledge bases such as KGs and data lakes support holistic integration for multimodal data arriving from heterogeneous sources, including tabular, key-value pairs, text, images, and multimedia data. Vector embedding represents each predicate and entity from diverse sources as a low-dimensional vector, such that the original structures and relations in the knowledge base are approximately preserved. Querying these vectors are essential for a wide range of applications, e.g., QA and semantic search. Finally, LLM pipelines are generally faster than traditional ML lifecycles – thanks to simpler prompt-based interactions without any requirement of re-training, making it easy to build AI pipelines around LLMs. Thus, the unification of three modern technologies LLMs, KGs, and vector DBs seem indispensable. There also remain fundamental challenges, e.g., how to create a holistic embedding across multiple modalities? It remains a nontrivial task to explain the results of LLMs and to incorporate domain knowledge – KGs could assist in both following graph RAG approaches. Analogously, adding human-in-the-loop and analyzing utility vs. privacy, bias, and fairness to derive quality solutions are important.

7. ACKNOWLEDGMENT

Khan acknowledges support from the Novo Nordisk Foundation grant NNF 22OC0072415. Ke is supported by Zhejiang Province's "Lingyan" R&D Project under Grant No. 2024C01259 and Yongjiang Talent Introduction Programme (2022A-237-G). Wu is supported by NSF under CNS-1932574, CNS-2028748, and OAC-2104007.

8. REFERENCES

[1] H. Abdallah and E. Mansour. Towards A GML-enabled Knowledge Graph Platform. In *ICDE*, 2023.
[2] M. Ali, M. Berrendorf, C. T. Hoyt, L. Vermue, M. Galkin, S. Sharifzadeh, A. Fischer, V. Tresp, and J. Lehmann. Bringing Light Into the Dark: A Large-Scale Evaluation of Knowledge Graph Embedding Models Under a Unified Framework. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):8825–8845, 2022.
[3] A. Andoni, P. Indyk, T. Laarhoven, I. Razenshteyn, and L. Schmidt. Practical and Optimal LSH for Angular Distance. In *NeurIPS*, 2015.
[4] A. Andoni, P. Indyk, and I. Razenshteyn. Approximate Nearest Neighbor Search in High Dimensions. In *International Congress of Mathematicians: Rio de Janeiro*, 2018.
[5] A. Antelmi, G. Cordasco, M. Polato, V. Scarano, C. Spagnuolo, and D. Yang. A Survey on Hypergraph Representation Learning. *ACM Computing Surveys*, 56(1):1–38, 2023.
[6] A. Babenko and V. Lempitsky. The Inverted Multi-index. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 37(6):1247–1260, 2014.
[7] C. D. T. Barros, M. R. F. Mendonça, A. B. Vieira, and A. Ziviani. A Survey on Embedding Dynamic Graphs. *ACM Comput. Surv.*, 55(1), 2021.
[8] M. Bawa, T. Condie, and P. Ganesan. LSH Forest: Self-tuning Indexes for Similarity Search. In *WWW*, 2005.
[9] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. v. d. Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. d. L. Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Simon Osindero, K. Simonyan, J. W. Rae, E. Elsen, and L. Sifre. Improving Language Models by Retrieving from Trillions of Tokens. In *ICML*, 2022.
[10] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral Networks and Locally Connected Networks on Graphs. In *ICLR*, 2014.
[11] H. Cai, V. W. Zheng, and K. C. Chang. A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. *IEEE Trans. Knowl. Data Eng.*, 30(9):1616–1637, 2018.
[12] L. Cai, J. Li, J. Wang, and S. Ji. Line Graph Neural Networks for Link Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):5103–5113, 2022.
[13] C. Chai, N. Tang, J. Fan, and Y. Luo. Demystifying Artificial Intelligence for Data Preparation. In *SIGMOD*, 2023.
[14] C. Chai, J. Wang, Y. Luo, Z. Niu, and G. Li. Data Management for Machine Learning: A Survey. *IEEE Trans. Knowl. Data Eng.*, 35(5):4646–4667, 2023.
[15] M. Chen, Z. Wei, B. Ding, Y. Li, Y. Yuan, X. Du, and J. Wen. Scalable Graph Neural Networks via Bidirectional Propagation. In *NeurIPS*, 2020.
[16] T. Chen, D. Qiu, Y. Wu, A. Khan, X. Ke, and Y. Gao. View-based Explanations for Graph Neural Networks. *Proc. ACM Manag. Data*, 2(1):40:1–40:27, 2024.
[17] P. Cui, X. Wang, J. Pei, and W. Zhu. A Survey on Network Embedding. *IEEE Trans. Knowl. Data Eng.*, 31(5):833–852, 2019.

[18] E. Dai and S. Wang. Towards Self-Explainable Graph Neural Network. In *CIKM*, 2021.

[19] A. Datta, M. Fredrikson, K. Leino, K. Lu, S. Sen, and Z. Wang. Machine Learning Explainability and Robustness: Connected at the Hip. In *KDD*, 2021.

[20] W. Dong, C. Moses, and K. Li. Efficient k-nearest Neighbor Graph Construction for Generic Similarity Measures. In *WWW*, 2011.

[21] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *NeurIPS*, 2015.

[22] K. Echihabi, T. Palpanas, and K. Zoumpatianos. New Trends in High-D Vector Similarity Search: AI-driven, Progressive, and Distributed. *PVLDB*, 14(12):3198–3201, 2021.

[23] L. Faber, A. K. Moghaddam, and R. Wattenhofer. When Comparing to Ground Truth is Wrong: On Evaluating GNN Explanation Methods. In *KDD*, 2021.

[24] W. Fan, Z. Fan, C. Tian, and X. L. Dong. Keys for Graphs. *PVLDB*, 8(12):1590–1601, 2015.

[25] W. Fan, R. Jin, M. Liu, P. Lu, C. Tian, and J. Zhou. Capturing Associations in Graphs. *PVLDB*, 13(11):1863–1876, 2020.

[26] W. Fan and P. Lu. Dependencies for Graphs. *TODS*, 44(2):5:1–5:40, 2019.

[27] P. Fang, A. Khan, S. Luo, F. Wang, D. Feng, Z. Li, W. Yin, and Y. Cao. Distributed Graph Embedding with Information-Oriented Random Walks. *PVLDB*, 16(7):1643–1656, 2023.

[28] C. Fu, C. Xiang, C. Wang, and D. Cai. Fast Approximate Nearest Neighbor Search with the Navigating Spreading-out Graph. *PVLDB*, 12:461—474, 2019.

[29] T. Funke, M. Khosla, M. Rathee, and A. Anand. Zorro: Valid, Sparse, and Stable Explanations in Graph Neural Networks. *IEEE Trans. Knowl. Data Eng.*, 35(8):8687–8698, 2023.

[30] J. Gao and C. Long. High-Dimensional Approximate Nearest Neighbor Search: with Reliable and Efficient Distance Comparison Operations. *PACMMOD*, 1(2):1–27, 2023.

[31] S. Gollapudi, N. Karia, V. Sivashankar, R. Krishnaswamy, N. Begwani, S. Raz, Y. Lin, Y. Zhang, N. Mahapatro, P. Srinivasan, et al. Filtered-DiskANN: Graph Algorithms for Approximate Nearest Neighbor Search with Filters. In *WWW*, 2023.

[32] A. Grover and J. Leskovec. Node2vec: Scalable Feature Learning for Networks. In *KDD*, 2016.

[33] S. Guan, H. Ma, P. Lin, and Y. Wu. GEDet: Adversarially Learned Few-shot Detection of Erroneous Nodes in Graphs. In *IEEE BigData*, 2020.

[34] S. Guan, H. Ma, M. Wang, and Y. Wu. GALE: Active Adversarial Learning for Erroneous Node Detection in Graphs. In *ICDE*, 2023.

[35] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh. WTF: The Who to Follow Service at Twitter. In *WWW*, 2013.

[36] W. L. Hamilton, Z. Ying, and J. Leskovec. Inductive Representation Learning on Large Graphs. In *NeurIPS*, 2017.

[37] X. Han, Z. Jiang, N. Liu, and X. Hu. G-Mixup: Graph Data Augmentation for Graph Classification. In *ICML*, 2022.

[38] X. He, Y. Tian, Y. Sun, N. V. Chawla, T. Laurent, Y. LeCun, X. Bresson, and B. Hooi. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. *arXiv preprint arXiv:2402.07630*, 2024.

[39] S. Horchidan and P. Carbone. ORB: Empowering Graph Queries through Inference. In *Joint Proceedings of the ESWC 2023 Workshops and Tutorials*, 2023.

[40] J. Hu, R. Cheng, Z. Huang, Y. Fang, and S. Luo. On Embedding Uncertain Graphs. In *CIKM*, 2017.

[41] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data. In *CIKM*, 2013.

[42] X. Huang, J. Zhang, D. Li, and P. Li. Knowledge Graph Embedding Based Question Answering. In *WSDM*, 2019.

[43] M. Hulsebos, X. Deng, H. Sun, and P. Papotti. Models and Practice of Neural Table Representations. In *SIGMOD*, 2023.

[44] J. Jang, H. Choi, H. Bae, S. Lee, M. Kwon, and M. Jung. CXL-ANNS: Software-Hardware Collaborative Memory Disaggregation and Computation for Billion-Scale Approximate Nearest Neighbor Search. In *USENIX ATC*, 2023.

[45] S. Jayaram Subramanya, F. Devvrit, H. V. Simhadri, R. Krishnawamy, and R. Kadekodi. Diskann: Fast Accurate Billion-point Nearest Neighbor Search on a Single Node. *NeurIPS*, 32, 2019.

[46] H. Jegou, M. Douze, and C. Schmid. Product Quantization for Nearest Neighbor Search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):117–128, 2010.

[47] B. Jin, G. Liu, C. Han, M. Jiang, H. Ji, and J. Han. Large Language Models on Graphs: A Comprehensive Survey. *IEEE Trans. Knowl. Data Eng.*, 36(12):8622–8642, 2024.

[48] W. Jin, L. Zhao, S. Zhang, Y. Liu, J. Tang, and N. Shah. Graph Condensation for Graph Neural Networks. In *ICLR*, 2022.

[49] P. Juillard, A. Bonifati, and A. Maur. Interactive Graph Repairs for Neighborhood Constraints. In *EDBT*, 2024.

[50] J. Kakkad, J. Jannu, K. Sharma, C. C. Aggarwal, and S. Medya. A Survey on Explainability of Graph Neural Networks. *IEEE Data Eng. Bull.*, 46(2):35–63, 2023.

[51] O. Keivani and K. Sinha. Improved Nearest Neighbor Search Using Auxiliary Information and Priority Functions. In *ICML*, 2018.

[52] A. Khan. Knowledge Graphs Querying. *SIGMOD Rec.*, 52(2):18–29, 2023.

[53] A. Khan and E. B. Mobaraki. Interpretability Methods for Graph Neural Networks. In *DSAA*, 2023.

[54] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis. Generalization through Memorization: Nearest Neighbor Language Models. In *ICLR*, 2020.

[55] B. Kim and F. Doshi-Velez. Machine Learning Techniques for Accountability. *AI Mag.*, 42(1):47–52, 2021.

[56] T. N. Kipf and M. Welling. Variational Graph Auto-Encoders. In *NeurIPS Workshop on Bayesian Deep Learning*, 2016.

[57] T. N. Kipf and M. Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*, 2017.

[58] J. Klicpera, A. Bojchevski, and S. Günnemann. Predict then Propagate: Graph Neural Networks meet Personalized PageRank. In *ICLR*, 2019.

[59] A. Kumar, M. Boehm, and J. Yang. Data Management in Machine Learning: Challenges, Techniques, and Systems. In *SIGMOD*, 2017.

[60] D. Le, K. Zhao, M. Wang, and Y. Wu. GraphLingo: Domain Knowledge Exploration by Synchronizing Knowledge Graphs and Large Language Models. In *ICDE*, 2024.

[61] A. Lerer, L. Wu, J. Shen, T. Lacroix, L. Wehrstedt, A. Bose, and A. Peysakhovich. Pytorch-BigGraph: A Large Scale Graph Embedding System. In *MLSys*, 2019.

[62] B. Li, W. Wang, Y. Sun, L. Zhang, M. A. Ali, and Y. Wang. GraphER: Token-Centric Entity Resolution with Graph Convolutional Neural Networks. In *AAAI*, 2020.

[63] C. Li, M. Zhang, D. G. Andersen, and Y. He. Improving Approximate Nearest Neighbor Search through Learned Adaptive Early Termination. In *SIGMOD*, 2020.

[64] W. Li, Y. Zhang, Y. Sun, W. Wang, M. Li, W. Zhang, and X. Lin. Approximate Nearest Neighbor Search on High Dimensional Data—Experiments, Analyses, and Improvement. *IEEE Trans. Knowl. Data Eng.*, 32(8):1475–1488, 2019.

[65] Y. Li, Z. Li, P. Wang, J. Li, X. Sun, H. Cheng, and J. X. Yu. A Survey of Graph Meets Large Language Model: Progress and Future Directions. In *IJCAI*, 2024.

[66] Y. Li, V. Zakhozhyi, D. Zhu, and L. J. Salazar. Domain Specific Knowledge Graphs as a Service to the Public. In *KDD*, 2020.

[67] H. Lin, M. Yan, X. Ye, D. Fan, S. Pan, W. Chen, and Y. Xie. A Comprehensive Survey on Distributed Training of Graph

Neural Networks. *Proc. IEEE*, 111(12):1572–1606, 2023.
[68] P. Lin, Q. Song, Y. Wu, and J. Pi. Repairing Entities using Star Constraints in Multirelational Graphs. In *ICDE*, 2020.
[69] X. Lin, Y. Peng, B. Choi, and J. Xu. Human-Powered Data Cleaning for Probabilistic Reachability Queries on Uncertain Graphs. *IEEE Trans. Knowl. Data Eng.*, 29(7):1452–1465, 2017.
[70] Z. C. Lipton. The Mythos of Model Interpretability. *Commun. ACM*, 61(10), 2018.
[71] C. Liu, H. Liu, H. Jin, X. Liao, Y. Zhang, Z. Duan, J. Xu, and H. Li. ReGNN: A ReRAM-Based Heterogeneous Architecture for General Graph Neural Networks. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, 2022.
[72] J. Liu, C. Yang, Z. Lu, J. Chen, Y. Li, M. Zhang, T. Bai, Y. Fang, L. Sun, P. S. Yu, and C. Shi. Towards Graph Foundation Models: A Survey and Beyond. *CoRR*, abs/2310.11829, 2023.
[73] S. Liu, Z. Zeng, L. Chen, A. Ainihaer, A. Ramasami, S. Chen, Y. Xu, M. Wu, and J. Wang. TigerVector: Supporting Vector Search in Graph Databases for Advanced RAGs. *CoRR*, 2501.11216, 2025.
[74] K. Lu, M. Kudo, C. Xiao, and Y. Ishikawa. HVS: Hierarchical Graph Structure Based on Voronoi Diagrams for Solving Approximate Nearest Neighbor Search. *PVLDB*, 15(2):246–258, 2021.
[75] K. Lu, C. Xiao, and Y. Ishikawa. Probabilistic Routing for Graph-Based Approximate Nearest Neighbor Search. In *ICML*, 2024.
[76] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang. Parameterized Explainer for Graph Neural Network. In *NeurIPS*, 2020.
[77] Y. Ma and J. Tang. *Deep Learning on Graphs*. Cambridge University Press, 2021.
[78] Y. A. Malkov and D. A. Yashunin. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(4):824–836, 2018.
[79] M. D. Manohar, Z. Shen, G. Blelloch, L. Dhulipala, Y. Gu, H. V. Simhadri, and Y. Sun. ParlayANN: Scalable and Deterministic Parallel Graph-Based Approximate Nearest Neighbor Search Algorithms. In *PPoPP24*, 2024.
[80] C. Mavromatis and G. Karypis. GNN-RAG: Graph Neural Retrieval for Large Language Model Reasoning. *CoRR*, 2024.
[81] E. Min, R. Chen, Y. Bian, T. Xu, K. Zhao, W. Huang, P. Zhao, J. Huang, S. Ananiadou, and Y. Rong. Transformer for Graphs: An Overview from Architecture Perspective. *CoRR*, abs/2202.08455, 2022.
[82] J. Mohoney, R. Waleffe, H. Xu, T. Rekatsinas, and S. Venkataraman. Marius: Learning Massive Graph Embeddings on a Single Machine. In *OSDI*, 2021.
[83] H. Ootomo, A. Naruse, C. Nolet, R. Wang, T. Feher, and Y. Wang. CAGRA: Highly Parallel Graph Construction and Approximate Nearest Neighbor Search for GPUs. In *ICDE*, 2024.
[84] J. J. Pan, J. Wang, and G. Li. Survey of Vector Database Management Systems. *VLDB J.*, 33(5):1591–1615, 2024.
[85] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Trans. Knowl. Data Eng.*, 36(7):3580–3599, 2024.
[86] R. D. Pasquale and S. Represa. Empowering Domain-Specific Language Models with Graph-Oriented Databases: A Paradigm Shift in Performance and Model Maintenance. *CoRR*, abs/2410.03867, 2024.
[87] L. Patel, P. Kraft, C. Guestrin, and M. Zaharia. ACORN: Performant and Predicate-Agnostic Search Over Vector Embeddings and Structured Data. In *SIGMOD*, 2024.
[88] Y. Peng, B. Choi, T. N. Chan, J. Yang, and J. Xu. Efficient Approximate Nearest Neighbor Search in Multi-dimensional Databases. *PACMMOD*, 1(1):1–27, 2023.
[89] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich. Data Management Challenges in Production Machine Learning. In *SIGMOD*, 2017.
[90] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann. Explainability Methods for Graph Convolutional Neural Networks. In *CVPR*, 2019.
[91] R. Pradhan, A. Lahiri, S. Galhotra, and B. Salimi. Explainable AI: Foundations, Applications, Opportunities for Data Management Research. In *ICDE*, 2022.
[92] D. Qiu, M. Wang, A. Khan, and Y. Wu. Generating Robust Counterfactual Witnesses for Graph Neural Networks. In *ICDE*, 2024.
[93] J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang. Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and Node2vec. In *WSDM*, 2018.
[94] A. Quamar, V. Efthymiou, C. Lei, and F. Özcan. Natural Language Interfaces to Data. *Found. Trends Databases*, 11(4):319–414, 2022.
[95] K. Rabbani, M. Lissandrini, and K. Hose. Shactor: improving the quality of large-scale knowledge graphs with validating shapes. In *Companion of the International Conference on Management of Data (SIGMOD)*, pages 151–154, 2023.
[96] M. Rathee, T. Funke, A. Anand, and M. Khosla. Bagel: A Benchmark for Assessing Graph Neural Network Explanations. *CoRR*, abs/2206.13983, 2022.
[97] H. Ren, M. Galkin, M. Cochez, Z. Zhu, and J. Leskovec. Neural Graph Reasoning: Complex Logical Query Answering Meets Graph Databases. *CoRR*, abs/2303.14617, 2023.
[98] H. Ren, W. Hu, and J. Leskovec. Query2box: Reasoning over Knowledge Graphs in Vector Space Using Box Embeddings. In *ICLR*, 2020.
[99] J. Ren, M. Zhang, and D. Li. HM-ANN: efficient billion-point nearest neighbor search on heterogeneous memory. In *NeurIPS*, 2020.
[100] A. Saxena, A. Tripathi, and P. P. Talukdar. Improving Multi-hop Question Answering over Knowledge Graphs using Knowledge Base Embeddings. In *ACL*, 2020.
[101] M. S. Schlichtkrull, N. D. Cao, and I. Titov. Interpreting Graph Neural Networks for NLP with Differentiable Edge Masking. In *ICLR*, 2021.
[102] M. S. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling. Modeling Relational Data with Graph Convolutional Networks. In *ESWC*, 2018.
[103] T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K. T. Schütt, K. Müller, and G. Montavon. Higher-Order Explanations of Graph Neural Networks via Relevant Walks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7581–7596, 2022.
[104] Y. Shao, H. Li, X. Gu, H. Yin, Y. Li, X. Miao, W. Zhang, B. Cui, and L. Chen. Distributed Graph Neural Network Training: A Survey. *ACM Comput. Surv.*, 56(8):191:1–191:39, 2024.
[105] J. Si, X. Gan, T. Xiao, B. Yang, D. Dong, and Z. Pang. STEGNN: Spatial-Temporal Embedding Graph Neural Networks for Road Network Forecasting. In *ICPADS*, 2022.
[106] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, H. Shum, and J. Guo. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model with Knowledge Graph. In *ICLR*, 2024.
[107] L. Sun, L. He, Z. Huang, B. Cao, C. Xia, X. Wei, and P. S. Yu. Joint Embedding of Meta-Path and Meta-Graph for Heterogeneous Information Networks. In *ICBK*, 2018.
[108] L. Sun, Z. Tao, Y. Li, and H. Arakawa. ODA: Observation-Driven Agent for integrating LLMs and Knowledge Graphs. In *ACL*, 2024.
[109] S. Suresh, P. Li, C. Hao, and J. Neville. Adversarial Graph Augmentation to Improve Graph Contrastive Learning. In *NeurIPS*, 2021.
[110] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li, and Y. Zhang. Learning and Evaluating Graph Neural Network Explanations Based on Counterfactual and Factual Reasoning. In *Web Conference*, 2022.

[111] D. Tang, J. Wang, R. Chen, L. Wang, W. Yu, J. Zhou, and K. Li. XGNN: Boosting Multi-GPU GNN Training via Global GNN MemoryStore. *PVLDB*, 17(5):1105–1118, 2024.

[112] J. Tang, Y. Yang, W. Wei, L. Shi, L. Su, S. Cheng, D. Yin, and C. Huang. GraphGPT: Graph Instruction Tuning for Large Language Models. In *SIGIR*, pages 491–500, 2024.

[113] Y. Tian. The world of graph databases from an industry perspective. *SIGMOD Rec.*, 51(4):60–67, 2022.

[114] Y. Tian, X. Zhao, and X. Zhou. DB-LSH 2.0: Locality-Sensitive Hashing With Query-Based Dynamic Bucketing. *IEEE Trans. Knowl. Data Eng.*, 2023.

[115] M. Vasimuddin, S. Misra, G. Ma, R. Mohanty, E. Georganas, A. Heinecke, D. D. Kalamkar, N. K. Ahmed, and S. Avancha. Distgnn: Scalable distributed training for large-scale graph neural networks. In *SC*, 2021.

[116] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph Attention Networks. In *ICLR*, 2018.

[117] P. Velickovic, R. Ying, M. Padovano, R. Hadsell, and C. Blundell. Neural Execution of Graph Algorithms. In *ICLR*, 2020.

[118] M. N. Vu and M. T. Thai. PGM-Explainer: Probabilistic Graphical Model Explanations for Graph Neural Networks. In *NeurIPS*, 2020.

[119] H. Wang, J. Wang, J. Wang, M. Zhao, W. Zhang, F. Zhang, X. Xie, and M. Guo. GraphGAN: Graph Representation Learning With Generative Adversarial Nets. In *AAAI*, 2018.

[120] J. Wang, P. Huang, H. Zhao, Z. Zhang, B. Zhao, and D. L. Lee. Billion-scale Commodity Embedding for E-commerce Recommendation in Alibaba. In *KDD*, 2018.

[121] J. Wang, X. Yi, R. Guo, H. Jin, P. Xu, S. Li, X. Wang, X. Guo, C. Li, X. Xu, et al. Milvus: A Purpose-built Vector Data Management System. In *SIGMOD*, 2021.

[122] M. Wang, X. Ke, X. Xu, L. Chen, Y. Gao, et al. MUST: An Effective and Scalable Framework for Multimodal Search of Target Modality. In *ICDE*, 2024.

[123] M. Wang, L. Lv, X. Xu, Y. Wang, Q. Yue, and J. Ni. An Efficient and Robust Framework for Approximate Nearest Neighbor Search with Attribute Constraint. In *NeurIPS*, 2023.

[124] M. Wang, H. Ma, A. Daundkar, S. Guan, Y. Bian, A. Sehirlioglu, and Y. Wu. CRUX: Crowdsourced Materials Science Resource and Workflow Exploration. In *CIKM*, 2022.

[125] M. Wang, H. Wu, X. Ke, Y. Gao, X. Xu, and L. Chen. An Interactive Multi-modal Query Answering System with Retrieval-Augmented Large Language Models. *PVLDB*, 17(12):1643–1656, 2024.

[126] M. Wang, W. Xu, X. Yi, S. Wu, Z. Peng, X. Ke, Y. Gao, X. Xu, R. Guo, and C. Xie. Starling: An I/O-Efficient Disk-Resident Graph Index Framework for High-Dimensional Vector Similarity Search on Data Segment. In *SIGMOD*, 2024.

[127] M. Wang, X. Xu, Q. Yue, and Y. Wang. A Comprehensive Survey and Experimental Comparison of Graph-based Approximate Nearest Neighbor Search. *PVLDB*, 14(11):1964—-1978, 2021.

[128] Y. Wang, A. Khan, T. Wu, J. Jin, and H. Yan. Semantic Guided and Response Times Bounded Top-k Similarity Search over Knowledge Graphs. In *ICDE*, 2020.

[129] Y. Wang, A. Khan, X. Xu, J. Jin, Q. Hong, and T. Fu. Aggregate Queries on Knowledge Graphs: Fast Approximation with Semantic-aware Sampling. In *ICDE*, 2022.

[130] Y. Wang, N. Lipka, R. A. Rossi, A. F. Siu, R. Zhang, and T. Derr. Knowledge Graph Prompting for Multi-Document Question Answering. In *AAAI*, 2024.

[131] C. Wei, B. Wu, S. Wang, R. Lou, C. Zhan, F. Li, and Y. Cai. Analyticdb-v: A Hybrid Analytical Engine Towards Query Fusion for Structured and Unstructured Data. *PVLDB*, 13(12):3152–3165, 2020.

[132] G. Weikum, X. L. Dong, S. Razniewski, and F. M. Suchanek. Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases. *Found. Trends Databases*, 10(2-4):108–490, 2021.

[133] Y. Wu, N. Hu, S. Bi, G. Qi, J. Ren, A. Xie, and W. Song. Retrieve-Rewrite-Answer: A KG-to-Text Enhanced LLMs Framework for Knowledge Graph Question Answering. *CoRR*, abs/2309.11206, 2023.

[134] Y. Wu, K. Ma, Z. Cai, T. Jin, B. Li, C. Zheng, J. Cheng, and F. Yu. Seastar: Vertex-centric Programming for Graph Neural Networks. In *EuroSys*, 2021.

[135] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Networks Learn. Syst.*, 2020.

[136] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Networks Learn. Syst.*, 32(1):4–24, 2021.

[137] F. Xia, K. Sun, S. Yu, A. Aziz, L. Wan, S. Pan, and H. Liu. Graph Learning: A Survey. *IEEE Trans. Artif. Intell.*, 2(2):109–127, 2021.

[138] J. Xia, H. Lin, Y. Xu, C. Tan, L. Wu, S. Li, and S. Z. Li. GNN Cleaner: Label Cleaner for Graph Structured Data. *IEEE Trans. Knowl. Data Eng.*, 2023.

[139] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How Powerful are Graph Neural Networks? In *ICLR*, 2019.

[140] Z. Xu, M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, and Z. Li. Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering. In *SIGIR*, 2024.

[141] R. Yang, J. Shi, X. Xiao, Y. Yang, J. Liu, and S. S. Bhowmick. Scaling Attributed Network Embedding to Massive Graphs. *PVLDB*, 14(1):37–49, 2020.

[142] M. Yasunaga, A. Bosselut, H. Ren, X. Zhang, C. D. Manning, P. Liang, and J. Leskovec. Deep Bidirectional Language-Knowledge Graph Pretraining. In *NeurIPS*, 2022.

[143] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In *NAACL-HLT*, 2021.

[144] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. GNNExplainer: Generating Explanations for Graph Neural Networks. In *NeurIPS*, 2019.

[145] D. Yu, C. Zhu, Y. Yang, and M. Zeng. JAKET: Joint Pre-training of Knowledge Graph and Language Understanding. In *AAAI*, 2022.

[146] Y. Yu, D. Wen, Y. Zhang, L. Qin, W. Zhang, and X. Lin. GPU-accelerated Proximity Graph Approximate Nearest Neighbor Search and Construction. In *ICDE*, 2022.

[147] H. Yuan, J. Tang, X. Hu, and S. Ji. XGNN: Towards Model-level Explanations of Graph Neural Networks. In *KDD*, 2020.

[148] H. Yuan, H. Yu, S. Gui, and S. Ji. Explainability in Graph Neural Networks: A Taxonomic Survey. *TPAMI*, 45(5):5782–5799, 2023.

[149] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji. On Explainability of Graph Neural Networks via Subgraph Explorations. In *ICML*, 2021.

[150] S. Yuan, X. Wu, and Y. Xiang. SNE: Signed Network Embedding. In *PAKDD*, 2017.

[151] Q. Yue, X. Xu, Y. Wang, Y. Tao, and X. Luo. Routing-Guided Learned Product Quantization for Graph-Based Approximate Nearest Neighbor Search. In *ICDE*, 2024.

[152] Q. Zhang, S. Xu, Q. Chen, G. Sui, J. Xie, Z. Cai, Y. Chen, Y. He, Y. Yang, F. Yang, et al. VBASE: Unifying Online Vector Similarity Search and Relational Queries via Relaxed Monotonicity. In *OSDI*, 2023.

[153] X. Zhang, A. Bosselut, M. Yasunaga, H. Ren, P. Liang, C. D. Manning, and J. Leskovec. GreaseLM: Graph REASoning Enhanced Language Models for Question Answering. In *ICLR*, 2022.

[154] Y. Zhang, H. Zhu, Z. Song, P. Koniusz, and I. King. COSTA: Covariance-Preserving Feature Augmentation for Graph Contrastive Learning. In *KDD*, 2022.

[155] Z. Zhang, P. Cui, and W. Zhu. Deep Learning on Graphs: A Survey. *IEEE Trans. Knowl. Data Eng.*, 34(1):249–270, 2022.

[156] Z. Zhang, Q. Liu, H. Wang, C. Lu, and C. Lee. ProtGNN: Towards Self-Explaining Graph Neural Networks. In *AAAI*, 2022.
[157] J. Zhao, Y. Dong, M. Ding, E. Kharlamov, and J. Tang. Adaptive Diffusion in Graph Neural Networks. In *NeurIPS*, 2021.
[158] K. Zhao, J. X. Yu, H. Zhang, Q. Li, and Y. Rong. A Learned Sketch for Subgraph Counting. In *SIGMOD*, 2021.
[159] T. Zhao, G. Liu, S. Günnemann, and M. Jiang. Graph Data Augmentation for Graph Machine Learning: A Survey. *IEEE Data Eng. Bull.*, 46(2):140–165, 2023.
[160] T. Zhao, Y. Liu, L. Neves, O. J. Woodford, M. Jiang, and N. Shah. Data Augmentation for Graph Neural Networks. In *AAAI*, 2021.
[161] T. Zhao, X. Tang, D. Zhang, H. Jiang, N. Rao, Y. Song, P. Agrawal, K. Subbian, B. Yin, and M. Jiang. AutoGDA: Automated Graph Data Augmentation for Node Classification. In *Learning on Graphs Conference*, 2022.
[162] T. Zhao, X. Zhang, and S. Wang. GraphSMOTE: Imbalanced Node Classification on Graphs with Graph Neural Networks. In *WSDM*, 2021.
[163] W. Zhao, S. Tan, and P. Li. Song: Approximate Nearest Neighbor Search on GPU. In *ICDE*, 2020.
[164] X. Zhao, Y. Tian, K. Huang, B. Zheng, and X. Zhou. Towards Efficient Index Construction and Approximate Nearest Neighbor Search in High-Dimensional Spaces. *PVLDB*, 16(8):1979–1991, 2023.
[165] C. Zheng, B. Zong, W. Cheng, D. Song, J. Ni, W. Yu, H. Chen, and W. Wang. Robust Graph Representation Learning via Neural Sparsification. In *ICML*, 2020.
[166] D. Zheng, C. Ma, M. Wang, J. Zhou, Q. Su, X. Song, Q. Gan, Z. Zhang, and G. Karypis. DistDGL: Distributed Graph Neural Network Training for Billion-Scale Graphs. In *IA3*, 2020.
[167] X. Zheng, Y. Liu, Z. Bao, M. Fang, X. Hu, A. W. Liew, and S. Pan. Towards Data-centric Graph Machine Learning: Review and Outlook. *CoRR*, abs/2309.10979, 2023.
[168] R. Zhu, K. Zhao, H. Yang, W. Lin, C. Zhou, B. Ai, Y. Li, and J. Zhou. Aligraph: A comprehensive graph neural network platform. *Proc. VLDB Endow.*, 12(12):2094–2105, 2019.
[169] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang. Graph Contrastive Learning with Adaptive Augmentation. In *WWW*, 2021.
[170] Z. Zhu, S. Xu, J. Tang, and M. Qu. GraphVite: A High-Performance CPU-GPU Hybrid System for Node Embedding. In *WWW*, 2019.
[171] C. Zuo, M. Qiao, W. Zhou, F. Li, and D. Deng. SeRF: Segment Graph for Range-Filtering Approximate Nearest Neighbor Search. In *SIGMOD*, 2024.

Reminiscences on Influential Papers

This issue's contributors cover papers that focus on different aspects of accessing *data*: RDFs, approximate nearest neighbor search, and compact hash tables. Furthermore, they all highlight the impact of the papers not only on their own research and career but also for the community in general. Enjoy reading!

While I will keep inviting members of the data management community, and neighboring communities, to contribute to this column, I also welcome unsolicited contributions. Please contact me if you are interested.

Pınar Tözün, *editor*
IT University of Copenhagen, Denmark
pito@itu.dk

Zoi Kaoudi
IT University of Copenhagen, Denmark
zoka@itu.dk

Thomas Neumann, Gerhard Weikum.
RDF-3X: a RISC-style engine for RDF.
In Proceedings of the VLDB Endowment, Volume 1, Issue 1, pages 647-659, 2008.

At the time that this paper [7] was published, I was doing my PhD and I was (feeling) part of the Semantic Web community. My topic was on distributed RDF query processing, optimization, and reasoning, so the paper was quite relevant to my work. Back then, there had been several proposals on different centralized RDF stores mostly from the Semantic Web community. Then, the RDF-3X paper appeared at VLDB and it really made a difference for me (and probably others). Its solution seemed very simple and elegant and its performance on runtime and scalability was remarkable. Although I continued my PhD by publishing in Semantic Web conferences, this paper, somehow subconsiously, played a role into deciding to change my career after my PhD and become part of the database community.

The paper proposed RDF-3X, an efficient and scalable RDF store engine. In contrast to most of the works so far that were proposing to store triples by splitting them in one table per property, it proposed to store the triples into a single gigantic 3-column table. Importantly, RDF-3X used all possible column permultations as indices together with dictionary encoding, and compression. This idea is so simple yet so powerful that made an impression on me. Later on, when I had already joined the database community, I also heard that the proposal of exhaustive indexing led to many interesting and controversial discussions among database researchers!

In the paper, the authors showed that having all these indices stored not only improves data access but makes the entire query processing faster. RDF-3X query processor followed a RISC-style design philosophy: it relied mostly on merge joins over sorted index lists. This was made possible thanks to its exhaustive indexing scheme. RDF-3X also encompassed a cost-based query optimizer to determine the right join ordering. It achieved very efficient cardinality estimation by utilizing the extensive indices and by maintaining an additional set of RDF-specific statistics for joins. The evaluation results showed huge performance benefits over simply loading the RDF data into a column-based or a row-based relational database. Interestingly, none of the open-source systems provided by the Semantic Web community could scale to the dataset sizes used in the experiments.

To conclude, I believe this paper has been very influencial across two different research communities (Semantic Web and databases) thanks to its simple idea and effective results. The fact that the system was also available to use gave the opportunity to

many researchers to come up with multiple follow-up works. I find it very inspiring to see simple ideas having such large impact in our research communities. I hope everyone takes that into account when reviewing papers nowadays.

Fatemeh Nargesian
University of Rochester, NY, USA
fnargesian@rochester.edu

Aristides Gionis, Piotr Indyk, Rajeev Motwani.
Similarity Search in High Dimensions via Hashing.
In Proceedings of the International Conference on Very Large Data Bases, pages 518-529, 1999.

When I saw Pınar's email asking for a couple of paragraphs for this column, one clear choice came to mind - and it remained unchanged despite all my procrastination. This paper [3] belongs to the line of work on Approximate Nearest Neighbor (ANN) Search.

The paper introduces the Locality-Sensitive Hashing (LSH) technique for approximate similarity search in high-dimensional spaces. The motivation is the inefficiency of traditional nearest neighbor search methods as dimensionality increases - a phenomenon known as the curse of dimensionality. The paper defines a family of hash functions for Euclidean distance such that the probability of collision is much higher for closer items than for others. These hash functions help build small-footprint signatures, consisting of hash values, that are similarity-preserving. This property enables the efficient retrieval of similar items without exhaustive comparisons. LSH achieves this by mapping fragments of signatures to buckets using standard hash functions. During query time, only the subset of buckets that have the same signatures as the query are searched. This drastically reduces computation cost. The paper provides provable guarantees on query time and approximation quality, specifically targeting (r, c·r)-nearest neighbor search (i.e., returning a point within distance c·r if any point exists within r).

I used various nearest neighbor search algorithms for my PhD research to overcome the scalability and efficiency challenges of data search in open repositories. While I always found these algorithms neat and useful, it was not until the thesis and paper-writing phase that I truly began to appreciate the elegance of LSH techniques. And, it was not until I started teaching the fundamentals of ANN search to my students that I came to appreciate the knitty gritty details of this paper and its previous and follow-up work around it.

Following the body of work that has built upon and around this paper has taught me a mindset and connected me to a broader world of practical research problems. While the VLDB'99 paper focuses on Euclidean space, the family of LSH has been developed for a variety of similarity measures: Cosine, Dot Product, etc. In recent years, indexes for approximate nearest neighbor search have regained popularity due to vector databases. More recent ANN indexes are shown to be more efficient and scalable in practice. Graph-based techniques, such as HNSW [6], NSG [2], and DiskANN [5], achieve search times of (poly/)logarithmic complexity by building proximity graphs with long-range and short-range links. The inverted index-based techniques such as FAISS, build an inverted index on the centroid of data partitions and only search for nearest neighbors within the partition associated with the closest centroid. Some of these techniques are in-memory and some are disk-based; some use product quantization and compression to reduce memory footprint; and, almost all scale to benchmark datasets of billions of points and are deployed in industry applications. The main difference, however, between the LSH family and the new generation of ANN, as shown by recent studies [4], is in their worst-case performance. The VLDB'99 paper shows that LSH has truly sublinear search time dependence on data size. Whereas, almost all others, with the exception of DiskANN, suffer from worst-case linear search time. Even DiskANN, which offers an improved worst-case complexity, does so at the cost of very slow preprocessing.

In the era of transformers and super-fast search techniques and all that give us great results and spark our curiosity to ask "why does it work and when not?", this paper has remained, for me, an example to follow in my own research; a reminder for when settling on an approximation for time-tradeoff, think about the guarantees of how fast and approximate our approximation is.

Niv Dayan
University of Toronto, Canada
nivdayan@cs.toronto.edu

John G. Cleary.
Compact hash tables using bidirectional linear probing.

In IEEE Transactions on Computers, Volume C-33, Issue 9, pages 828-834, 1984.

My first encounter with Bloom filters was a love at first sight. A Bloom filter is a space-efficient probabilistic data structure that allows you to test whether a key is definitely not in a set, or possibly is. By compressing a set of keys into a compact bit array in memory, Bloom filters enable fast membership tests that help avoid expensive storage or network lookups when the key in question is absent.

Originally proposed in 1970, Bloom filters have become a mainstay in modern systems. Yet, despite their widespread use, they suffer from two important limitations. First, they do not support deletions or dynamic expansion as the dataset grows. Second, they only support point queries-checking for the presence of individual keys-but not range queries, which are essential in many database applications that need to determine whether an entire range is empty.

Over the past few years, our lab has been exploring ways to overcome these limitations. A key source of inspiration in our journey has been the paper "Compact Hash Tables Using Bidirectional Linear Probing" by John G. Cleary from 1984 [1]. This paper presents a compact hash table design built on four major ideas:

1. Quotienting – storing only the suffix of a key, inferring the prefix from its location.

2. Robin Hood Hashing – resolving hash collisions by searching sequentially for a nearby available slot, while keeping colliding entries adajcent.

3. Dual bitmaps – marking the start and end of clustered entries that map to the same slot.

4. Prefix sum arrays – aggregating the 1s in the bitmaps to support efficient navigation and search.

I appreciate this paper not only for its technical contributions, but also for its clarity—it explains complex concepts in intuitive, problem-driven terms without compromising on rigor. Many filter data structures developed over the past 15 years have drawn from this framework to provide more memory-efficient alternatives to Bloom filters that also support deletions.

These ideas have directly informed our research. In our InfiniFilter (SIGMOD 2023) and Aleph Filter (VLDB 2024) papers, we designed filters that can dynamically expand while maintaining a stable false positive rate. We achieve this by using variable-sized fingerprints padded with unary codes. Supporting this feature required a hash table whose collision resolution does not depend on the fingerprints themselves, unlike in Cuckoo filters. The Cleary data structure was an ideal fit.

Our recent work on range filters has also benefited from this foundation. Memento Filter (SIGMOD 2025), for instance, stores variable-length payloads alongside keys—something that's naturally supported by Robin Hood hashing, where entries can simply be pushed and pulled in sequence. Our newest range filter, Diva (currently under submission to VLDB 2025), goes a step further by encoding key infixes and relying on the Cleary structure being order-preserving. Interestingly, the Cleary data structure wasn't originally proposed with these use-cases in mind, yet it turns out to be exactly what we needed for each of the above projects.

We would also like to acknowledge the influential paper "A General-Purpose Counting Filter: Making Every Bit Count" by Pandey et al. from SIGMOD 2017 [8]. This work demonstrated how to search similar structures using rank and select primitives implemented efficiently using CPU instructions introduced in Intel's Haswell line of processors (Bit Manipulation Instruction Set 2). It also showed how to succinctly encode counters to allow representing multisets. We built on top of this excellent design and its codebase in many of our projects.

Of course, these are just a few highlights among the many foundational papers that have shaped our thinking. Our pursuit may be compact data structures, but the foundation beneath them is anything but small.

1. REFERENCES

[1] J.G. Clerry. Compact Hash Tables Using Bidirectional Linear Probing. *IEEE Transactions on Computers*, C-33(9):828–834, 1984.

[2] Cong Fu, Chao Xiang, Changxu Wang, and Deng Cai. Fast approximate nearest neighbor search with the navigating spreading-out graph. *Proc. VLDB Endow.*, 12(5):461–474, 2019.

[3] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity Search in High Dimensions via Hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases*, page 518–529, 1999.

[4] Piotr Indyk and Haike Xu. Worst-case Performance of Popular Approximate Nearest Neighbor Search Implementations: Guarantees and Limitations. In *Proceedings of the 37th International Conference on Neural*

Information Processing Systems, NIPS '23, 2023.

[5] Ravishankar Krishnaswamy, Magdalen Dobson Manohar, and Harsha Vardhan Simhadri. The DiskANN library: Graph-Based Indices for Fast, Fresh and Filtered Vector Search. *IEEE Data Eng. Bull.*, 48(3):20–42, 2024.

[6] Yury A. Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. *CoRR*, abs/1603.09320, 2016.

[7] Thomas Neumann and Gerhard Weikum. RDF-3X: a RISC-style engine for RDF. *Proc. VLDB Endow.*, 1(1):647–659, 2008.

[8] Prashant Pandey, Michael A. Bender, Rob Johnson, and Rob Patro. A General-Purpose Counting Filter: Making Every Bit Count. In *Proceedings of the 2017 ACM International Conference on Management of Data*, page 775–787, 2017.

On nurturing doubt and intuition

Sihem AMER-YAHIA , CNRS, Univ. Grenoble Alpes, France

When I was 5 growing up in Algiers, my parents took me to the music conservatory and asked me which instrument I wanted to learn to play. I did not know the answer and I said I wanted to sign up for Ballet dancing. Since then, dancing has been central to my life. When they asked me what I wanted to do after high school, I did not know the answer and I chose computer science because I heard my Math teacher say it was the future. When my husband asked me to marry him, I literally answered "I am hungry, let's get dinner". Since then, dinner has been a special moment for us. I was in my mid-career transition when I moved from NYC to Barcelona to Doha and then to Grenoble. I love not knowing the answer and yet making great choices. My mid-career advice is: nurture doubt, develop intuition, and learn to make great choices.

The biggest change you will experience when entering your mid-career phase is a widening of your choices. That applies to your collaborators, your academic responsibilities, the conferences you will attend, the projects you will get involved in as a leader or as a partner, the people you want to mentor, the life choices you get to make, the grants you will apply for, the topics you want to work on, the services you get to complete for your research community, and the students and collaborators you will interact with on a daily basis. Choice is a blessing and a responsibility.

On choice and responsibility

I don't know for you but when I turned 40, about a decade ago, I was suddenly faced with a diverse set of choices: whether to go to Academia or Industry, the country and even the continent to live in and work on, my life partner, and many other important decisions, such as whether I should continue with classical Ballet dancing or switch to softer modern Jazz. I feel like mid-career transitions are just like that. At this stage in your career, you have the luxury of choice. With that, comes responsibility. All of a sudden, you become a role model. That shift from being a junior-on-the way-to-senior is so palpable you cannot ignore it. Expect to be the senior in the room and when you are, ask yourself what piece of advice you can provide to others. Are you the role model you want to be? What is the image you are projecting? Are you being nice to juniors? How can you help them?

Surround yourself with seniors

A senior-to-be often attracts juniors in search of role models. Being a role model for those juniors is a major endeavor you should pursue. At the same time, you need to find energy and inspiration in others. It helps to be the junior in the room for that. Actively search out mentors to learn from, nurture your relationships and enjoy learning from them. I find that as I entered my mid-career phase, my discussions with seniors became more open and more in touch with my feelings. I was feeling less concerned about what they could think about me and I could more easily express my opinions. My interactions with seniors became more fruitful from then on. Also, remember there will always be smarter people around you. Stay humble, listen to others, and do not forget you can always learn from them, be they seniors, mid-careers, or juniors.

Don't be afraid to play senior

Playing senior is not easy for a soon-to-become-senior. Seniors know that and they also know that if you're trying hard and if you believe in what you're doing, their role is to help. People who are more senior than you can read you. They have been there and they can see bits of themselves in you. So, if you want to play senior, don't shy away from asking them for advice.

All of a sudden, you are the center of attention

In my experience, the mid-career phase is when you get approached by most people: juniors who need support, seniors who need your expertise, and value your energy, and other seniors-to-be who know you will complement their expertise, colleagues who need Associate Editors and Editors in Chief, those who seek keynote speakers, those looking for support letters, those searching members of their hiring committee, those looking for

reviewers of PhD theses, those seeking to nominate not-so-junior and not-senior researchers for awards. That, added to the fact that you are expected to take on new responsibilities such as project leadership and coordination, will require you to rethink time management. Simple quantification measures could be applied. Quantify the effort you make for each task and aim for balance. Think about your more junior colleagues who are seldomly invited to be conference officers and suggest they get invited instead of you.

Remain motivated

Depending on where you live you may become tenured as soon as you are hired (it is the case in Academia in France), never (it's the case in Academia in Chile), or just now. If you were in France or in Chile you'd have asked yourself the motivation question already. In other places where you just got tenure, you need to ask yourself that question because and assuming it was until now, getting tenure is not your motivation anymore. What motivates humans? Peer recognition, monetary compensation, altruism, challenges, pride. Ask yourself what motivates you and how you can keep going. Freedom to explore anything is something that has always motivated me.

Remember where you came from and pay attention at your surroundings

It is helpful to remember where you came from and what your progression has been and what it took to get there. Do you want to continue with the same levers? Do you want to make the same sacrifices? And while you're at it, remember those who lifted you, those you are lifting, and those who help you do your work. Undeniably, some seniors will be jealous because you are still young and vibrant and some juniors may try to ride your wave with little effort on their part. Learn to cut off toxic relationships and move on. I do not have a recipe for that and I fell into some hurtful traps. I would have paid more attention if I had known.

It is okay to fail, try understanding why

You just succeeded in achieving a major step in your professional life. Now, you can afford to fail from time to time. Failure is a gift and an opportunity to learn from. Remember we are competing with very smart and hardworking folks, and remember that real life is not always fair. Talk to others, explain your failures, and ask them their opinion on why some project, research idea, paper, application of yours did not make it. We do not talk enough about our failures and we can learn so much from them. One of my favorite events is the Failed Aspirations in Database Systems (FADS@VLDB) workshop. I hope that as a community we could hold more events like those. I have been working with people in other disciplines, medical doctors, economists, law professors, and education scientists. It is only after I talked to some of them in more relaxed social settings, over lunch or dinner, that I understood why some of our attempts failed: why did they not promptly share data with me after promising to do so? Why could I not get them to contribute a paragraph or two when I needed that? Why is our student feeling frustrated? Why are we not able to converge toward the same goal? More freedom undeniably leads to more failures. All you need is to learn to deal with your failures.

Understand what success means to you

Now that you are a junior senior, you need to think about what you are seeking next. If you made it this far it means your community recognized you for some work and can associate your name to some research topics you helped further. You can now work on your right to be "forgotten" for that and remembered 5, 10 years from now, for something else. That other thing is a combination of research and service. It needs to be a new research topic because you want to keep innovating. I feel very proud of making that shift and bringing social computing to the database community and I thank TCDE and VLDB for providing me awards for that. It also needs to be about service because your community would benefit from what you can bring (thank you Tamer for running the advice to mid-careers). I feel very proud of having succeeded to establish the diversity equity and inclusion initiative in the database community and I thank SIGMOD for recognizing my efforts with an award.

Your network

As a junior senior, you have probably figured that collaborating with others is essential in research. Let me argue why it matters even more now. I was told very early on in my career that once you find a collaborator

who matches your interests, make sure to nurture that collaboration. I have been told to learn to see the best side in others. I have been lucky to meet amazing colleagues. I also believe the harder I work the luckier I get. All that forms a magic recipe for fruitful collaborations. I firmly believe that being kind to others and being generous with one's time and effort pays off. So, my strongest advice is to strive to be kind and generous because the more senior you are, which is undeniably the path you are set to follow, the scarier you become to the younger generation. So, be both nice and firm.

Understand your recovery activities and remember to pause and smell the flowers

How do you recover from effort and refuel for the next steps? Hobbies, exercise, family, friends, travels, doing nothing, drinking tea, gardening, cooking for hours? You may want to think about your current recovery activities and if you want to take on new ones. Just like starting new research topics, starting new recovery activities is a renewal.

Prepare to face slower times

All researchers experience a slowdown some time in their career. While that may bring some frustration, think about how to exploit that moment. Time to consider different publication venues? Time to attend conferences from other communities? Time to venture into new topics? Time to focus more on listening to your students and mentees?

Prepare your future by being attuned to your time

Every research community has its preferences and nurtures them. This will continue to be the case and it will help you ride the wave of hot research. As a mid-career researcher you can start asking yourself what will make your research community still relevant in 5, 10 years and what will make it societally relevant. I find this exercise difficult and rewarding. I started working on ranking algorithms that accounted for relevance and diversity because I wondered how we could make Boolean database queries more relevant to people. I started working on algorithmic fairness on labor platforms because I asked myself the question of how people were treated on crowdsourcing platforms. I learned how to deploy principled user studies when I paid attention to closing the loop between experiments and algorithm design. I started working on Education because I wondered if people learned anything by collaborating with others in solving tasks. All these questions led me to expanding my horizons to other research communities in Computer Science, SIGIR, TWC, and ICWSM, but also other sciences, Law, Economics, Education, and Medicine.

Today, one may ask how their work may contribute to inequality reduction, climate protection, and quality education. While making that effort you need to confront and reject the feeling that you have to be doing something big to be doing anything at all. What are the smallest steps you could take today, and with whom, to make a difference. How many other jobs provide one the opportunity to ask themselves such a question? You have the best job in the world, a perfect balance of doubt, intuition, and continuous learning and intellectual effort. Remember that.

A few months ago, I was approached by colleagues from our AI institute in Grenoble who work on creative thinking and creative design. They suggested we put together a project on how data and generative AI could impact human creativity. Engaging myself in such a project appeared to me as the paroxysm of doubt. At the same time, intuitively, it felt like a natural next step in my research: after treating humans as mere receivers of query results, data producers in online platforms, workers on crowdsourcing and labor marketplaces, learners on an online education platform, here I am asking myself how to help people be more creative. If I had allowed myself to think about this rationally, I would not have accepted. For the first time I am going to start my research by running qualitative and quantitative experiments with human subjects to gather their interactions with Generative AI as a companion or as a tool, before seeking to solve any technical question. I feel very excited about that despite not knowing where I am heading and not even knowing if it is a great choice. One thing I know this time is that the journey will be rewarding and that is something we have the luxury to afford in our job.

The Case for a New Cloud-Native Programming Model with Pure Functions

Ana Klimovic
ETH Zurich
aklimovic@ethz.ch

Cloud evolution: Over the past two decades, the cloud has become the dominant platform for running all kinds of applications, from data analytics to web services. In the process, cloud platforms have evolved from renting virtual machines (VMs) on-demand to offering elastic compute and storage services. While the ability to support legacy applications was critical in the early days of cloud to ease migration from on-premise, today's users commonly develop *cloud-native* applications by composing cloud storage services (e.g., S3), compute services (e.g., AWS Lambda), data analytics services (e.g., BigQuery), machine learning services (Azure ML), and elastic databases (e.g., Snowflake [4]). With this approach, users no longer need to explicitly provision CPU/memory/storage for their applications, as the elastic services automatically scale-out based on load and bill users for the resources consumed [7].

Opportunity and obstacle: By abstracting resource management from users, elastic cloud services have the potential to optimize resource allocation, task scheduling, and data movement under the hood to improve overall performance and energy-efficiency. Multi-tenant cloud services like AWS S3 and Lambda can optimize resource allocation with a global view across users [8].

However, a major optimization obstacle is that today's cloud programming model captures very little about the resource requirements and data access patterns of individual applications, leaving cloud services with little information to apply optimizations. Despite new cloud-native models like Functions as a Service (FaaS), today's cloud is still built around the principle of executing *opaque*[1] user applications inside VMs. For example, FaaS platforms execute a user function as an opaque unit in a MicroVM [1]. Each serverless function arbitrarily combines custom computation logic and calls to external cloud services for data passing. The platform is not aware of inter-function nor inter-service dependencies, making it difficult to optimize task scheduling and data prefetching. As a result, serverless functions often spend a large fraction of their execution time blocked on I/O [5]. To avoid idling CPU cores while functions block, the platform can multiplex many VMs per core. However, context switching securely between VMs adds latency [2] and comes with a high memory footprint, as the platform must allocate the total memory needed for all in-flight VMs.

Rethink the programming model: A promising way to enable cloud platforms to improve performance and resource efficiency is to rethink the cloud-native programming model, such that users develop applications in ways that provide the cloud platform with key information to guide task scheduling and data prefetching optimizations.

We propose a programming model that strictly separates compute tasks (custom user logic) and I/O tasks (interactions between cloud services). In this new paradigm, users express applications by composing two types of functions: 1) *pure compute functions*, i.e., untrusted user code snippets that compute exclusively on declared inputs and produce declared outputs and 2) *I/O functions*, i.e., trusted code implemented by the platform and exposed to users as a library, enabling interaction with other services, like storage.

Separating compute and I/O has several benefits. First, it makes application dataflow explicit to the platform, enabling data prefetching and task scheduling optimizations [3, 11]. For example, the platform can colocate functions that need to exchange data and allocate CPU cores and memory to functions only when their inputs are ready. Second, separating I/O tasks (which require interaction with the operating system and hence have a large attack surface) from other user code enables executing user code with more lightweight isolation mechanisms than canonical VMs [9, 10] to improve performance. Finally, separating computation and I/O in the programming model simplifies offloading each type of task to hardware accelerators, as accelerators are typically specialized for either fast computation or fast I/O. We are currently exploring these ideas in *Dandelion* [6], a new serverless platform.

[1] *Opaque* execution refers to execution with no awareness of application characteristics, such as data dependencies.

References

[1] Alexandru Agache, Marc Brooker, Alexandra Iordache, Anthony Liguori, Rolf Neugebauer, Phil Piwonka, and Diana-Maria Popa. "Firecracker: Lightweight Virtualization for Serverless Applications". In: *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*. 2020.

[2] Amazon Web Services. *The Security Design of the AWS Nitro System*. https://docs.aws.amazon.com/whitepapers/latest/security-design-of-aws-nitro-system/security-design-of-aws-nitro-system.html. 2022.

[3] Ankit Bhardwaj, Meghana Gupta, and Ryan Stutsman. "On the Impact of Isolation Costs on Locality-aware Cloud Scheduling". In: *12th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 20)*. 2020.

[4] Benoit Dageville, Thierry Cruanes, Marcin Zukowski, Vadim Antonov, Artin Avanes, Jon Bock, Jonathan Claybaugh, Daniel Engovatov, Martin Hentschel, Jiansheng Huang, Allison W. Lee, Ashish Motivala, Abdul Q. Munir, Steven Pelley, Peter Povinec, Greg Rahn, Spyridon Triantafyllis, and Philipp Unterbrunner. "The Snowflake Elastic Data Warehouse". In: *Proceedings of the 2016 International Conference on Management of Data*. SIGMOD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 215–226. ISBN: 9781450335317. DOI: 10.1145/2882903.2903741. URL: https://doi.org/10.1145/2882903.2903741.

[5] Yuhan Deng, Angela Montemayor, Amit Levy, and Keith Winstein. "Computation-Centric Networking". In: *Proceedings of the 21st ACM Workshop on Hot Topics in Networks*. HotNets '22. 2022.

[6] Tom Kuchler, Michael Giardino, Timothy Roscoe, and Ana Klimovic. "Function as a Function". In: *Proceedings of the 2023 ACM Symposium on Cloud Computing*. SoCC '23. 2023.

[7] Ingo Müller, Renato Marroquín, and Gustavo Alonso. "Lambada: Interactive Data Analytics on Cold Data Using Serverless Cloud Infrastructure". In: *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. SIGMOD '20. Portland, OR, USA: Association for Computing Machinery, 2020, pp. 115–130. ISBN: 9781450367356. DOI: 10.1145/3318464.3389758. URL: https://doi.org/10.1145/3318464.3389758.

[8] Vivek Narasayya and Surajit Chaudhuri. "Multi-Tenant Cloud Data Services: State-of-the-Art, Challenges and Opportunities". In: *Proceedings of the 2022 International Conference on Management of Data*. SIGMOD '22. Philadelphia, PA, USA: Association for Computing Machinery, 2022, pp. 2465–2473. ISBN: 9781450392495. DOI: 10.1145/3514221.3522566. URL: https://doi.org/10.1145/3514221.3522566.

[9] Vasily A Sartakov, Lluís Vilanova, David Eyers, Takahiro Shinagawa, and Peter Pietzuch. "CAP-VMs: Capability-Based Isolation and Sharing for Microservices". In: *Proceedings of Operating Systems Design and Implementation*. OSDI '22. 2022.

[10] Simon Shillaker and Peter Pietzuch. "Faasm: Lightweight Isolation for Efficient Stateful Serverless Computing". In: *2020 USENIX Annual Technical Conference (USENIX ATC 20)*. 2020.

[11] Minchen Yu, Tingjia Cao, Wei Wang, and Ruichuan Chen. "Following the data, not the function: Rethinking function orchestration in serverless computing". In: *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. 2023, pp. 1489–1504.

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!.

LLM+KG@VLDB'24 Workshop Summary

Arijit Khan[1] Tianxing Wu[2] Xi Chen[3]
[1]Aalborg University, Denmark [2]Southeast University, China [3]Platform and Content Group, Tencent, China
[1]arijitk@cs.aau.dk [2]tianxingwu@seu.edu.cn [3]jasonxchen@tencent.com

ABSTRACT

The unification of large language models (LLMs) and knowledge graphs (KGs) has emerged as a hot topic. At the LLM+KG'24 workshop, co-located with VLDB 2024 in Guangzhou, China, the key theme explored was important data management challenges and opportunities due to the effective interaction between LLMs and KGs. The report outlines major directions and approaches presented by various speakers during the workshop.

1. INTRODUCTION

LLMs, a relatively newer form of generative AI, have become ubiquitous, revolutionizing natural language processing with applications ranging from solving problems, streamlining workflows, augmenting analytics, code synthesis, to accessing information via conversational functionality, e.g., Copilots and digital assistants. LLMs are skilled at learning stochastic language patterns as parametric knowledge, and thus predicting next tokens for the given contexts. However, LLMs may lack consistent knowledge representations. Hence, they experience hallucinations and generate unreliable or factually incorrect outputs. KGs can offer external, factual, and up-to-date knowledge to LLMs via, e.g., retrieval augmented methods, improving the LLMs' accuracy, consistency, and transparency. On the other hand, LLMs can also facilitate data curation, knowledge extraction, KG creation, completion, embedding, and various downstream tasks over KGs such as recommendation and question answering (QA). Furthermore, the unification of LLMs and KGs creates new data management opportunities and challenges in consistency, scalability, knowledge editing, privacy, fairness, explainability, data regulations, human-in-the-loop, software-hardware collaboration, cloud-based solutions, and AI-native databases.

The LLM+KG'24 ambition was to provide a unique platform to researchers and practitioners for presentation of the latest research results, new technology developments and applications, as well as outline the vision for next-generation solutions in the trending topic of unifying LLMs+KGs. The workshop also aims at discussing what interesting opportunities are awaiting for the data management researchers in this greener pasture.

The full-day workshop included 3 keynote talks on the synergies between LLMs and KGs, 1 industrial invited talk on GraphRAG [31], 9 peer-reviewed research papers from different countries in North and South America, Europe, Asia, and Africa, and a panel discussion on the unification of LLMs, KGs, and Vector databases (Vector DBs). The detailed program is available at [20].

2. KEYNOTES

The program featured three keynotes by Guilin Qi (Southeast University, China), Haofen Wang (Tongji University, China), and Wei Hu (Nanjing University, China).

2.1 Integrating KGs with LLMs: From the Perspective of Knowledge Engineering

The first keynote talk on integrating KGs with LLMs from the knowledge engineering point-of-view was given by Guilin Qi from Southeast University. Prof. Qi started with the enlightening question, *"What is knowledge?"* and shared a number of interesting perspectives. First, according to the Oxford Dictiory, knowledge is the information, understanding, and skills that one gains with education or experience. Second, informally speaking, knowledge can be fact-based, description of information (e.g., text, image), or skills obtained by practice. Third, one way to decide whether artificial intelligence (AI) has human intelligence or not could possibly be by the AI's ability to learn and apply knowledge. Fourth, a Knowledge Base (KB) is a collection of knowledge, including documents, images, triples, rules, parameters of neural networks, etc. Fifth, a KG is a data structure for representing knowledge using a graph. Prof. Qi further emphasized 'KGs as knowledge bases' as follows: "*Knowledge graphs originated from how machines represent knowledge, use graph structures to describe relationships between things, developed in the rise of Web technologies, and landed in applications such as search engine, intelligent QA, and recommender systems"*.

Next, Prof. Qi introduced the fundamentals of language models and whether they can be used as 'paramet-

ric knowledge bases' [36]. He compared the reasoning capabilities of LLMs and KGs, their advantages and disadvantages, and elaborated significant research scopes and practical values due to the complementary nature and mutual enhancements between symbolic knowledge of KGs and parametric knowledge of LLMs.

In the direction of 'KGs for LLMs', Prof. Qi discussed how KGs enhance pre-training [56], fine-tuning [51], inference [52], prompting [7, 23], retrieval/ knowledge augmented generation [54], knowledge editing [60, 55], knowledge fusion [39], and knowledge validation [15] of LLMs. In the other direction of 'LLMs for KGs', he mentioned knowledge engineering by LLMs, where LLMs can act as both resources (e.g., data augmentation) and enablers (e.g., encoding, reading comprehension, and QA). He also stated several opportunities such as LLMs for entity and relation extraction, triple generation, ontology matching [14], entity alignment [18], knowledge base QA [43], ontology reasoning [47], and KG reasoning [33, 58, 46], among others.

Prof. Qi concluded by underlining interesting opportunities due to LLM+KG integration and the engineering efforts required to work properly, e.g., OpenKG [4] and new knowledge platforms to support generalizable, trustable, and stable knowledge services. His concluding remark was to look at "*Language as the "form", knowledge as the "heart", and graph as the "skeleton"*."

2.2 Industry-level KG Platforms for Large-scale, Diverse, and Dynamic Scenarios

In the second keynote talk on industry-level knowledge graph platforms, Haofen Wang from Tongji University stated that traditional knowledge semantic frameworks such as RDF/OWL and labeled property graph (LPG) have major limitations in knowledge modeling and management and are often inadequate in modern business scenarios. Prof. Wang provided examples of the Ant Group KG applications in the finance sectors. First, the data sources for KGs have grown tremendously from text to heterogeneous enterprise data, e.g., semi-structured/ unstructured user-generated/ professionally generated contents, structured profiles from business operations, transactions, and logs, requiring to implement knowledge hierarchies and lightweight alignments of diverse sources through programmable methods. Second, knowledge representations have emerged from binary static structures to multi-dimensional dynamic associations in temporal and spatial dimensions, therefore deep collaborative information from multiple aspects of entities, events, concepts, contexts, etc. are required for real-world applications such as merchant management and risk control. In summary, the development of the KG technology does not match the expectations of the new paradigm of an industry-scale, unified, automated knowledge modeling framework for the entire life cycles of businesses, with the ability to evolve and support continuous business iterations.

Next, Prof. Wang introduced the Semantic-enhanced Programmable Graph framework (SPG) developed by the Ant Group and OpenKG [6] that integrates the structural aspects of LPG with the semantic nature of RDF – overcoming the semantic complexity of RDF/OWL, while also retaining the simplicity of LPG and its compatibility with the big data systems. The SPG layered architecture consists of several modules. (i) SPG-Schema is responsible for the schema design. (ii) SPG- Programming, a programmable framework, deals with knowledge construction, knowledge evolution, expert experience projection, and knowledge graph reasoning. (iii) SPG-Engine is responsible for the execution process of SPG syntax. (iv) SPG-Controller is the control center subsystem, taking care of the control framework, command distribution, and plugin integration. (v) SPG-LLM interacts with LLMs for natural language understanding.

Prof. Wang concluded by discussing the potentials of SPG and LLM-guided next-generation industry-level cognitive engines, as well as building an AI framework based on the OpenSPG knowledge engine.

2.3 KG-based LLM Fine-tuning

The third keynote talk on KG-enhanced LLM fine-tuning was given by Wei Hu from Nanjing University. Prof. Hu emphasized the knowledge gap problem of general-purpose LLMs – they often lack accurate domain knowledge, resulting in inaccurate and unreliable outputs, and even difficulty in real-world applications.

Among various knowledge enhancement techniques for LLMs, Prof. Hu focused on an LLM fine-tuning framework with adaptive integration of multi-source KGs, consisting of knowledge extraction, knowledge fusion, and KG-enhanced LLMs. In the field of knowledge extraction, he introduced problems such as domain named entity recognition, document-level relation extraction [49], continual event extraction [50], document-level event causality identification [27], and continual relation extraction [48]. In knowledge fusion, Prof. Hu discussed embedding-based entity alignment [44, 41, 12, 40], knowledge transfer [19, 53], adding human-in-the-loop [17, 16], benchmarking, and the OpenEA toolkit [42]. In KG-enhanced LLMs fine-tuning, he introduced KnowLA [28], a knowledgeable adaptation method for PEFT (parameter efficient fine-tuning), particularly for LoRA (Low-Rank Adaptation). (i) KnowLA with LoRA can align the space of the LLM with the space of KG embeddings, and (ii) KnowLA can activate the parameterized potential knowledge that originally exists in the LLM, even though the used KG does not contain such knowledge.

Prof. Hu concluded with interesting applications of KG-enhanced LLMs in translating configuration files dur-

ing device replacements in communication networks and unified PEFT+RAG (Retrieval-Augmented Generation).

3. INDUSTRIAL INVITED TALK

Siwei Gu and Yihang Yu from NebulaGraph [3] delivered an inspiring industrial talk on GraphRAG [31], i.e., **Integrating GenAI with Graph: Innovations and Insights from NebulaGraph**. RAG is a technique to optimize the output of an LLM so that it references an authoritative, up-to-date KB outside of its training corpus before generating a response. Given a user's query, the classic RAG approach uses vector similarity to retrieve semantically similar matches. It also builds offline indexes over embedding vectors to speed-up online retrieval, but partitioning knowledge across chunks can lose global context/ inter-relationships. Connection-oriented retrieval (e.g., join and multi-hop queries) as well as addressing broad, global questions that require synthesizing insights from the entire data can be challenging when the context is spread over multiple chunks.

To resolve the aforementioned issues, NebulaGraph launched industry-first GraphRAG [31] – a technology harnessing the power of knowledge graphs to provide retrieval methods with a more comprehensive contextual understanding and thereby assisting users in obtaining cost-effective, smarter, and more precise search results with an LLM. In particular, it uses a KG to model the external KB, shows the relationships between entities, which can more accurately understand the query intent, and then uses retrieval enhancement for LLMs. For instance, one can use graph reasoning or subgraph retrieval to find relevant contexts through relationships. Users can push domain Knowledge to KG schema and relationships [54]. Furthermore, one can apply graph-based indexing for a more comprehensive retrieval of context, since graph indexing helps in connecting fragmented knowledge.

Gu and Yu concluded by discussing potential directions about various indexing and retrieval strategies in graphRAG [1], node importance finding [13], chain-of-exploration [37], and query-focused summarization [10].

4. RESEARCH PAPERS

The peer-reviewed research papers presented in this workshop can be broadly classified into three categories.

4.1 LLMs for KGs

KGs are difficult to construct due to the high cost. KG querying is also challenging due to their incompleteness, users requiring to have full knowledge of the query language (e.g., SPARQL, Cypher), and the large and complex KG schema. LLMs can assist in KG construction via prompt engineering without huge labeling efforts, and improve the usability and performance of natural language QA with their strong understanding and generalization capabilities. Nie et al. leverage domain-specific knowledge from ontology and Chain-of-Thought prompts to extract higher-quality triples from unstructured text [32]. Groves et al. empirically compare in-context learning, fine-tuning, and supervised learning in automated knowledge curation for biomedical ontologies [11]. Mou et al. explore in-context learning capabilities of GPT-4 for instruction driven adaptive knowledge graph construction, while also proposing a self-reflection mechanism to enable LLMs to critically evaluate their outputs and learn from errors using examples [29]. Mustafa et al. use the W3C Open Digital Rights Language (ODRL) ontology and its documentation to formulate prompts in large language models and generate usage policies in ORDL from natural language instructions [30].

4.2 KGs for LLMs

LLMs hallucinate due to lack of context or knowledge gap. Offering domain-specific and up-to-date knowledge through KGs can enhance the accuracy, consistency, transparency, and the overall capabilities of LLMs. Liu et al. propose a collaborative LLMs method for open-set object recognition, incorporating KGs to alleviate hallucination of LLMs [26]. Wang et al. study a novel infuser-guided knowledge integration framework to integrate unknown knowledge into LLMs efficiently without unnecessary overlap of known knowledge [45].

4.3 Unifying LLMs+KGs

The third category of papers simultaneously leverage the factual knowledge of KGs and the parametric knowledge of LLMs to mutually enhance each other. Zhang et al. introduce OneEdit – a neural-symbolic prototype system for collaborative knowledge editing using natural language and facilitating easy-to-use knowledge management with KGs and LLMs [57]. Khorashadizadeh et al. present a survey on the synergy between LLMs and KGs [22]. Cavalleri et al. present the SPIREX system to extract triples from scientific literature involving RNA molecules [8]. They exploit schema constraints in the formulation of LLM prompts and also utilize graph machine learning on an RNA-based KG to assess the plausibility of extracted triples.

5. PANEL

The workshop was concluded with a panel discussion [9] on the unification of LLMs, KGs, and Vector databases (Vector DBs). The panelists were Wei Hu (Nanjing University, China), Shreya Shankar (UC Berkeley, USA), Haofen Wang (Tongji University, China), and Jianguo Wang (Purdue University, USA).

LLMs, KGs, and Vector DBs: Synergy and Opportunities for Data Management. The LLM+KG'24 chairs first asked some questions. **Q1.** What are the synergies among LLMs, Vector DBs, and graph data management including KGs? **Q2.** What are the roles of DBs in LLMs + KGs + Vector data management? **Q3.** How can LLMs + KGs + Vector data enhance data management? **Q4.** What are the significance of human-in-the loop and responsible AI in LLM systems and Vector DBs? How can KGs help in these aspects? **Q5.** How can academia + industry partnership and interdisciplinary collaborations advance this field? What would be the roles of benchmarking, open-source models, tools, and datasets?

The panelists added further perspectives to those questions. While some aspects of these technologies may seem part of the hype cycle, the foundational ideas behind the integration of LLMs, Vector DBs, and KGs are well-grounded in addressing real-world data challenges, and LLMs are definitely a key to genAI. They can reinforce each other by combining structured/ semi-structured and well-curated data for accuracy (e.g., KGs), efficient data retrieval (Vector DBs), and contextual understanding (LLMs), ensuring robust querying, reasoning, and interpretability. Many old DB ideas are relevant around LLMs' self-consistency, thinking step-by-step, etc. [34]. For the deploymentment of LLMs in data pipeline, bolt-on data quality constraints for LLM-generated data is crucial [38]. LLMs over graph-based applications need both vector- and graph-based RAGs, e.g., consider queries like *"What do others say about my papers?"* or *"Find competitors with similar products to mine and analyze their pricing strategies for different products"*. Relational DBs may support efficient vector data management [59], e.g., PASE is a highly optimized generalized vector database based on PostgreSQL.

These technologies will enhance databases, knowledge engineering, and data science by enabling more dynamic and responsive search and query responses, facilitating richer interactions with multi-modal data from diverse sources, integrating domain-specific understanding and learning deep semantics. Many potential areas or success stories include NLIDB (natural language interfaces for data bases)/ Text2SQL, query optimization, data curation, neural DBs, self-driving DBs, data education, OpenKG+SPG, and declarative systems for AI workloads (e.g., Palimpzest [24], LOTUS [35]). The synergy is particularly transformative in domains like personalized healthcare and financial analytics.

Transparency and explainability are key challenges in this domain. LLMs make mistakes and require guardrails. Both human-in-the-loop and KGs can align LLMs by providing contextual relevance, factual information, and feedback based on preferences. Ultimately, developing AI systems that adhere to ethical guidelines, emphasizing safety, accountability, fairness, privacy, and transparency is crucial for deploying them in the real world.

This is an interdisciplinary area, and the DB community is well-positioned to own the data pre-processing and validation parts of LLM pipelines [5]. However, encouraging idea exchanges by integrating expertises from fields like DB, ML, NLP, HCI, and CV can drive innovations and create end-to-end solutions/ systems. Fostering academia + industry partnerships would require aligning objectives, e.g., industries can offer internships and GPU resources, co-fund initiatives for practical impact and knowledge exchange, while also leading the LLM developments. Benchmarking and providing open source models, tools, and data are important to enhances accessibility, innovation, and community collaboration. Recently, there are also concerns, e.g., many benchmark datasets and empirical studies, domain-specific LLMs reporting only "biased" results, etc.

Finally, the panel concluded by discussing open problems such as conducting neural-symbolic reasoning, managing complex, dynamic KGs, scaling integration and reducing costs, guardrailing LLMs, ensuring data privacy and compliance, and various engineering challenges.

6. FUTURE DIRECTIONS

We conclude that there are several ongoing works in the area of LLMs+KGs, with many open problems, e.g.,

• **Integration of Vector and Graph Databases.** Leveraging vector DBs for GraphRAG creates new opportunities such as combining graph DBs with vector DBs [25], using graph DBs as semantic caches of LLMs enabling semantic matching for new graph queries instead of expensive LLM API calls [21], optimizing the index creation and similarity search over large-scale graph embeddings, and hardware acceleration.

• **Efficient and Explainable GraphRAG.** The efficiency of relevant subgraphs retrieval and raking is challenging in GraphRAG as KGs are large and the context length of LLM is limited. In GraphRAG, KGs can enhance explainability by linking LLM-generated answers to explicit KG relationships, while also acting as guardrails to validate answers against factual knowledge.

• **Knowledge Conflict and Dynamic Integration.** Aligning LLMs+KGs is a critical challenge in knowledge engineering since overlap and conflict occur when integrating new knowledge from external sources into LLMs. Incremental updates to KGs and dynamic integration with LLMs are crucial for up-to-date knowledge integration.

The second edition of the workshop LLM+Graph'25 [2] will be held in conjunction with VLDB 2025 with a broader perspective, since we shall focus on data management for the general topic of LLM+graph computing, rather than only data management for LLM+KG.

7. REFERENCES

[1] GraphRAG LlamaIndex Workshop. https://colab.research.google.com/drive/1tLjOg2ZQuIClfuWrAC2LdiZHCov8oUbs.

[2] LLM+Graph: The Second International Workshop on Data Management Opportunities in Bringing LLMS with Graph Data. https://seucoin.github.io/workshop/llmg2025/.

[3] NebulaGraph. https://github.com/vesoft-inc/nebula.

[4] OpenKG. https://github.com/OpenKG-ORG.

[5] *DEEM '24: Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning*. Association for Computing Machinery, 2024.

[6] Ant Group and OpenKG. Semantic-enhanced Programmable Knowledge Graph (SPG) White paper (v1.0). https://spg.openkg.cn/en-US, 2023.

[7] J. Baek, A. F. Aji, and A. Saffari. Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering. In *NLRSE@ACL*, 2023.

[8] E. Cavalleri, M. Soto-Gomez, A. Pashaeibarough, D. Malchiodi, H. Caufield, J. Reese, C. J. Mungall, P. N. Robinson, E. Casiraghi, G. Valentini, and M. Mesiti. SPIREX: Improving LLM-based Relation Extraction from RNA-focused Scientific Literature using Graph Machine Learning. In *Workshops at the International Conference on Very Large Data Bases (VLDB)*, 2024.

[9] X. Chen, W. Hu, A. Khan, S. Shankar, H. Wang, J. Wang, and T. Wu. Large Language Models, Knowledge Graphs, and Vector Databases: Synergy and Opportunities for Data Management (A Report on the LLM+KG@VLDB24 Workshop's Panel Discussion). https://wp.sigmod.org/?p=3813, 2024.

[10] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *CoRR*, abs/2404.16130, 2024.

[11] E. Groves, M. Wang, Y. Abdulle, H. Kunz, J. Hoelscher-Obermaier, R. Wu, and H. Wu. Benchmarking and Analyzing In-Context Learning, Fine-tuning and Supervised Learning for Biomedical Knowledge Curation: A Focused Study on Chemical Entities of Biological Interest. In *Workshops at the International Conference on Very Large Data Bases (VLDB)*, 2024.

[12] L. Guo, Q. Zhang, Z. Sun, M. Chen, W. Hu, and H. Chen. Understanding and Improving Knowledge Graph Embedding for Entity Alignment. In *ICML*, pages 8145–8156, 2022.

[13] B. J. Gutiérrez, Y. Shu, Y. Gu, M. Yasunaga, and Y. Su. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. *CoRR*, abs/2405.14831, 2024.

[14] S. Hertling and H. Paulheim. OLaLa: Ontology Matching with Large Language Models. In *K-CAP*, pages 131–139, 2023.

[15] N. Hu, J. Chen, Y. Wu, G. Qi, S. Bi, T. Wu, and J. Z. Pan. Benchmarking Large Language Models in Complex Question Answering Attribution using Knowledge Graphs. *CoRR*, abs/2401.14640, 2024.

[16] J. Huang, W. Hu, Z. Bao, Q. Chen, and Y. Qu. Deep Entity Matching with Adversarial Active Learning. *VLDB J.*, 32(1):229–255, 2023.

[17] J. Huang, Z. Sun, Q. Chen, X. Xu, W. Ren, and W. Hu. Deep Active Alignment of Knowledge Graph Entities and Schemata. *Proc. ACM Manag. Data*, 1(2):159:1–159:26, 2023.

[18] X. Jiang, Y. Shen, Z. Shi, C. Xu, W. Li, Z. Li, J. Guo, H. Shen, and Y. Wang. Unlocking the Power of Large Language Models for Entity Alignment. In *ACL*, pages 7566–7583, 2024.

[19] X. Jiao, W. Li, X. Wu, W. Hu, M. Li, J. Bian, S. Dai, X. Luo, M. Hu, Z. Huang, D. Feng, J. Yang, S. Feng, H. Xiong, D. Yu, S. Li, J. He, Y. Ma, and L. Liu. PGLBox: Multi-GPU Graph Learning Framework for Web-Scale Recommendation. In *KDD*, pages 4262–4272, 2023.

[20] A. Khan, T. Wu, and X. Chen. LLM+KG: Data Management Opportunities in Unifying Large Language Models + Knowledge Graphs. In *Workshops at the International Conference on Very Large Data Bases (VLDB)*, 2024.

[21] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations (ICLR)*, 2020.

[22] H. Khorashadizadeh, F. Z. Amara, M. Ezzabady, F. Ieng, S. Tiwari, N. Mihindukulasooriya, J. Groppe, S. Sahri, and F. Benamara. Research Trends for the Interplay between Large Language Models and Knowledge Graphs. In *Workshops at the International Conference on Very Large Data Bases (VLDB)*, 2024.

[23] S. Ko, H. Cho, H. Chae, J. Yeo, and D. Lee. Evidence-Focused Fact Summarization for Knowledge-Augmented Zero-Shot Question Answering. *CoRR*, abs/2403.02966, 2024.

[24] C. Liu, M. Russo, M. J. Cafarella, L. Cao, P. B. Chen, Z. Chen, M. J. Franklin, T. Kraska, S. Madden, and G. Vitagliano. A Declarative System for Optimizing AI Workloads. *CoRR*, abs/2405.14696, 2024.

[25] S. Liu, Z. Zeng, L. Chen, A. Ainihaer, A. Ramasami, S. Chen, Y. Xu, M. Wu, and J. Wang. TigerVector: Supporting Vector Search in Graph Databases for Advanced RAGs. *CoRR*, 2501.11216, 2025.

[26] X. Liu, Y. Wu, Y. Zhou, J. Chen, H. Wang, Y. Liu, and S. Wan. Enhancing Large Language Models with Multimodality and Knowledge Graphs for Hallucination-free Open-set Object Recognition. In *Workshops at the International Conference on Very Large Data Bases (VLDB)*, 2024.

[27] Y. Liu, X. Jiang, W. Zhao, W. Ge, and W. Hu. Dual Graph Convolutional Networks for Document-Level Event Causality Identification. In *APWeb-WAIM*, page 114–128, 2023.

[28] X. Luo, Z. Sun, J. Zhao, Z. Zhao, and W. Hu. KnowLA: Enhancing Parameter-efficient Finetuning with Knowledgeable Adaptation. In *NAACL*, pages 7153–7166, 2024.

[29] Y. Mou, L. Liu, S. Sowe, D. Collarana, and S. Decker. Leveraging LLMs Few-shot Learning to Improve Instruction-driven Knowledge Graph Construction. In *Workshops at the International Conference on Very Large Data Bases (VLDB)*, 2024.

[30] D. M. Mustafa, A. Nadgeri, D. Collarana, B. T. Arnold, C. Quix, C. Lange, and S. Decker. From Instructions to ODRL Usage Policies: An Ontology Guided Approach. In *Workshops at the International Conference on Very Large Data Bases (VLDB)*, 2024.

[31] NebulaGraph. NebulaGraph Launches Industry-First Graph RAG: Retrieval-Augmented Generation with LLM Based on Knowledge Graphs. https://www.nebula-graph.io/posts/graph-RAG, 2023.

[32] J. Nie, X. Hou, W. Song, X. Wang, X. Zhang, X. Jin, S. Zhang, and J. Shi. Knowledge Graph Efficient Construction: Embedding Chain-of-Thought into LLMs. In *Workshops at the International Conference on Very Large Data Bases (VLDB)*, 2024.

[33] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Trans. Knowl. Data Eng.*, 36(7):3580–3599, 2024.

[34] A. G. Parameswaran, S. Shankar, P. Asawa, N. Jain, and Y. Wang. Revisiting Prompt Engineering via Declarative Crowdsourcing. In *CIDR*, 2024.

[35] L. Patel, S. Jha, C. Guestrin, and M. Zaharia. LOTUS: Enabling Semantic Queries with LLMs Over Tables of Unstructured and Structured Data. *CoRR*, abs/2407.11418, 2024.

[36] F. Petroni, T. Rocktäschel, S. Riedel, P. S. H. Lewis, A. Bakhtin, Y. Wu, and A. H. Miller. Language Models as Knowledge Bases? In *EMNLP-IJCNLP*, pages 2463–2473, 2019.

[37] D. Sanmartin. KG-RAG: Bridging the Gap Between Knowledge and Creativity. *CoRR*, abs/2405.12035, 2024.

[38] S. Shankar, H. Li, P. Asawa, M. Hulsebos, Y. Lin, J. Zamfirescu-Pereira, H. Chase, W. Fu-Hinthorn, A. G. Parameswaran, and E. Wu. SPADE: Synthesizing Data Quality Assertions for Large Language Model Pipelines. *Proc. VLDB*

Endow., 2024.

[39] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, H. Shum, and J. Guo. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model with Knowledge Graph. *CoRR*, abs/2307.07697, 2023.

[40] Z. Sun, W. Hu, C. Wang, Y. Wang, and Y. Qu. Revisiting Embedding-Based Entity Alignment: A Robust and Adaptive Method. *IEEE Trans. Knowl. Data Eng.*, 35(8):8461–8475, 2023.

[41] Z. Sun, J. Huang, X. Xu, Q. Chen, W. Ren, and W. Hu. What Makes Entities Similar? A Similarity Flooding Perspective for Multi-sourced Knowledge Graph Embeddings. In *ICML*, pages 32875–32885, 2023.

[42] Z. Sun, Q. Zhang, W. Hu, C. Wang, M. Chen, F. Akrami, and C. Li. A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs. *Proc. VLDB Endow.*, 13(11):2326–2340, 2020.

[43] Y. Tan, D. Min, Y. Li, W. Li, N. Hu, Y. Chen, and G. Qi. Can ChatGPT Replace Traditional KBQA Models? An In-Depth Analysis of the Question Answering Performance of the GPT LLM Family. In *ISWC*, pages 348–367, 2023.

[44] X. Tian, Z. Sun, and W. Hu. Generating Explanations to Understand and Repair Embedding-Based Entity Alignment. In *ICDE*, pages 2205–2217, 2024.

[45] F. Wang, R. Bao, S. Wang, W. Yu, Y. Liu, W. Cheng, and H. Chen. InfuserKI: Enhancing Large Language Models with Knowledge Graphs via Infuser-Guided Knowledge Integration. In *Workshops at the International Conference on Very Large Data Bases (VLDB)*, 2024.

[46] J. Wang, T. Wu, S. Chen, Y. Liu, S. Zhu, W. Li, J. Xu, and G. Qi. unKR: A Python Library for Uncertain Knowledge Graph Reasoning by Representation Learning. In *SIGIR*, pages 2822–2826, 2024.

[47] K. Wang, G. Qi, J. Li, and S. Zhai. Can Large Language Models Understand DL-Lite Ontologies? An Empirical Study. *CoRR*, abs/2406.17532, 2024.

[48] X. Wang, Z. Wang, and W. Hu. Serial Contrastive Knowledge Distillation for Continual Few-shot Relation Extraction. In *ACL*, pages 12693–12706, 2023.

[49] X. Wang, Z. Wang, W. Sun, and W. Hu. Enhancing Document-Level Relation Extraction by Entity Knowledge Injection. In *ISWC*, pages 39–56, 2022.

[50] Z. Wang, X. Wang, and W. Hu. Continual Event Extraction with Semantic Confusion Rectification. In *EMNLP*, pages 11945–11955, 2023.

[51] Z. Wang, Q. Zhang, K. Ding, M. Qin, X. Zhuang, X. Li, and H. Chen. InstructProtein: Aligning Human and Protein Language via Knowledge Instruction. In *ACL*, pages 1114–1136, 2024.

[52] Y. Wei, Q. Huang, Y. Zhang, and J. T. Kwok. KICGPT: Large Language Model with Knowledge in Context for Knowledge Graph Completion. In *EMNLP Findings*, pages 8667–8683, 2023.

[53] H. Wu, X. Zhu, and W. Hu. A Blockchain System for Clustered Federated Learning with Peer-to-Peer Knowledge Transfer. *Proc. VLDB Endow.*, 17(5):966–979, 2024.

[54] Z. Xu, M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, and Z. Li. Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering. In *SIGIR*, pages 2905–2909, 2024.

[55] Y. Yao, P. Wang, B. Tian, S. Cheng, Z. Li, S. Deng, H. Chen, and N. Zhang. Editing Large Language Models: Problems, Methods, and Opportunities. In *EMNLP*, pages 10222–10240, 2023.

[56] D. Yu, C. Zhu, Y. Yang, and M. Zeng. JAKET: Joint Pre-training of Knowledge Graph and Language Understanding. In *AAAI*, pages 11630–11638, 2022.

[57] N. Zhang, Z. Xi, Y. Luo, P. Wang, B. Tian, Y. Yao, J. Zhang, S. Deng, M. Sun, L. Liang, Z. Zhang, X. Zhu, J. Zhou, and H. Chen. OneEdit: A Neural-Symbolic Collaboratively Knowledge Editing System. In *Workshops at the International Conference on Very Large Data Bases (VLDB)*, 2024.

[58] Y. Zhang, Z. Chen, L. Guo, Y. Xu, W. Zhang, and H. Chen. Making Large Language Models Perform Better in Knowledge Graph Completion. In *MM*, pages 233–242, 2024.

[59] Y. Zhang, S. Liu, and J. Wang. Are There Fundamental Limitations in Supporting Vector Data Management in Relational Databases? A Case Study of PostgreSQL. In *ICDE*, pages 3640–3653, 2024.

[60] C. Zheng, L. Li, Q. Dong, Y. Fan, Z. Wu, J. Xu, and B. Chang. Can We Edit Factual Knowledge by In-Context Learning? In *EMNLP*, pages 4862–4876, 2023.

Visual Analytics Challenges and Trends in the Age of AI: The BigVis Community Perspective

Nikos Bikakis*
Hellenic Mediterranean University &
Archimedes/Athena RC
Greece

Panos K. Chrysanthis
University of Pittsburgh
USA

Guoliang Li
Tsinghua University
China

George Papastefanatos
Athena RC
Greece

Lingyun Yu
Xi'an Jiaotong-Liverpool University
China

1 Introduction

This report provides insights into the challenges, emerging topics, and opportunities related to human–data interaction and visual analytics in the AI era.

The BigVis 2024[1] organizing committee conducted a survey among experts in the field. They invited the Program Committee members and the authors of accepted papers to share their views. *Thirty-two scientists* from diverse research communities, including Databases, Information Visualization, and Human–Computer Interaction, participated in the study. These scientists, representing both industry and academia, provided valuable insights into the current and future landscape of the field.

In this report, we analyze the survey responses and compare them to the findings of a similar study conducted four years ago [2]. The results reveal some interesting insights. First, *many of the critical challenges identified in the previous survey remain highly relevant today*, despite being unrelated to AI. Meanwhile, the field's landscape has significantly evolved, with *most of today's vital challenges not even being mentioned in the earlier survey*, underscoring the profound impact of AI-related advancements.

By summarizing the perspectives of the research community, this report aims to shed light on the key challenges, emerging trends, and potential research directions in human–data interaction and visual analytics in the AI era.

2 Survey Overview

The survey is divided into two parts. The first is related to challenges (Sec. 3), and the second focuses on emerging research topics (Sec. 4).

The participants were requested to answer six questions, either by filling out free-text fields or selecting from the options provided. The survey was anonymous, since the questions related to personal information are optional, e.g., name, county, affiliation. The survey required, on average, about three to five minutes to be completed.

Participants Demographics. We intended to find scientists from different research communities (e.g., Databases, Information Visualization, HCI), and from industry and academia. To this end, the survey was disseminated to the BigVis 2024 Program Committee members (58 members) and to the authors of accepted BigVis 2024 papers (34 authors). At the end, 32 *of the scientists invited completed the survey*.

The following *characteristics of the participants are collected* (Fig. 1):

- **Scientific Field** (Fig. 1a): The options were: (a) *Database*; (b) *Information Visualization*; (c) *Data Minning*; (d) *Human–Computer Interaction*; (e) *Computer Graphics*; and (f) *Other*. Most of the participants belong to Information Visualization (47%) and Database (37%) communities, while 16% belong to others research fields.
- **Career** (Fig. 1b): The options were: *Academic* (81%) and *Industry* (19%).
- **Position** (Fig. 1c): The options were: *Professor* (59%); *Researcher* (28%); and *Analyst/Scientist/ Engineer* (13%).

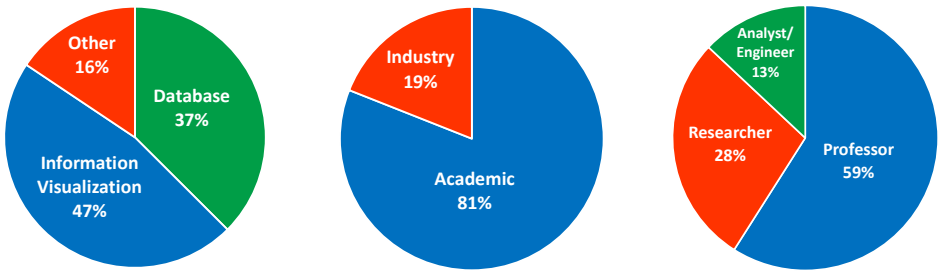
3 Research Challenges

In this section we outline the survey's results regarding research challenges related to data visualization and visual analytics.

In the first part (Sec. 3.1), the participants were asked to vote on today's importance of the challenges emerged four years ago in a 2020 report, titled "*Big*

*Corresponding author.

[1] 7th *Intl. Workshop on Big Data Visual Exploration & Analytics*, in conjunction with the 50th *Intl. Conf. on Very Large Databases* (VLDB 2024), Guangzhou, China. More details about the BigVis workshops can be found in [7].



(a) Scientific Field **(b) Career** **(c) Position**

Figure 1: Participants Demographics

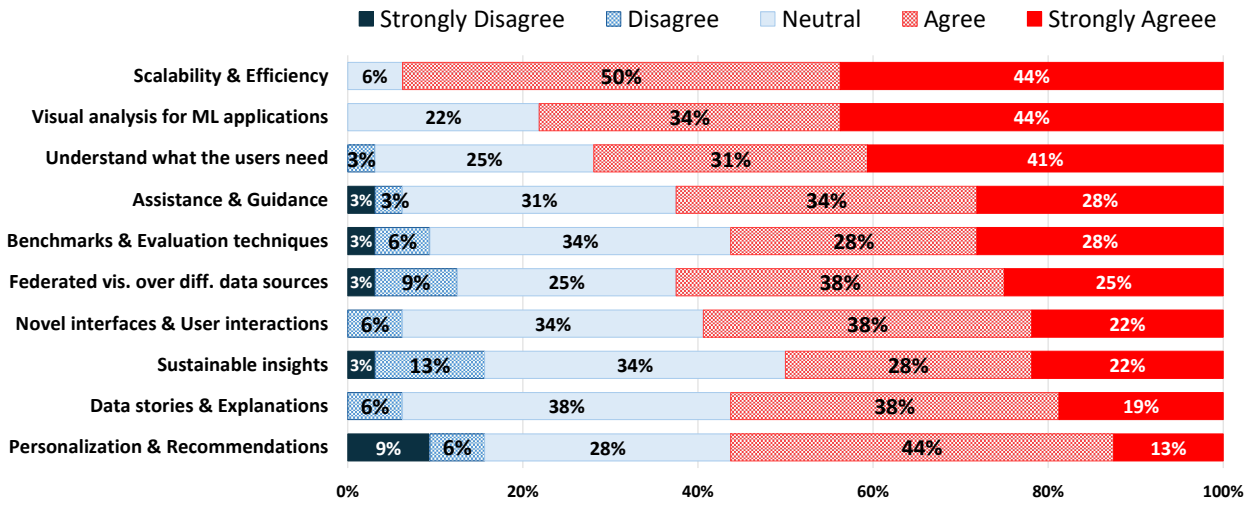


Figure 2: The Importance of the 2020 Challenges Today ["*Is this challenge important today*?"]

Data Visualization and Analytics: Future Research Challenges and Emerging Applications" [2]. In the next part (Sec. 3.2), participants suggested a challenge they consider the most important today, regardless of whether it was included in 2020 challenges.

3.1 The Challenges of the 2020 Report

This section presents the results of the survey regarding today's importance of the ten challenges identified four years ago in the 2020 report. Particularly, in the context of the 3rd International Workshop on Big Data Visual Exploration and Analytics (BigVis 2020), the organizing committee invited 14 *distinguished scientists*, from different communities to provide their insights regarding the *challenges and the applications they find more interesting in coming years, related to the areas of Big Data visualization and analytics*.

The challenges indicated in the 2020 report were: (a) *Support scalability & efficiency*; (b) *Enable visual analysis for ML applications*; (c) *Understand what the users need*; (d) *Build novel interfaces & user interactions*; (e) *Assistance & guidance*; (f) *Generate data stories & explanations*; (g) *Enable federated visualization over different data sources*; (h) *Develop benchmarks & evaluation techniques*; (i) *Provide sustainable insights*; and (k) *Enable personalization & recommendations*.

Question 1

The participants were asked to vote on the *ten challenges stated in the 2020 report, based on their importance/emerge*. Particularly, the participants rated each challenge using a five-level Likert scale (i.e., Strongly Disagree, Disagree, Neutral, Agree, and Strongly Agree) on the question "*Is this challenge important today*?".

Question 1 Responses. The results are presented in Figure 2 via a percent stacked bar chart. Participants voted "*Scalability & efficiency*" as the most important challenge, with 94% of participants indicating they agree or strongly agree, and 0% disagree or strongly disagree.

The second most significant is the "*Visual analysis for ML applications*" challenge, with 78% (resp. 0%) of the participants state that agree or strongly agree (resp. disagree or strongly disagree). "*Sustainable insights*" is the challenge in which most of the participants disagree (13%) or strongly disagree (3%). Finally, "*Personalization & recommendations*" is voted as the least important challenge.

It is worth mentioning that for *all challenges, at least* 50% *of the participants strongly agree or agree on today's importance of the challenge. Similar results can be observed when considering importance scores*.

3.2 The Challenges of the 2024 Survey

This section presents the challenges stated by the participants as the most important, regardless of whether they were included in 2020 challenges.

Question 2

The participants were asked to provide in a free text the *challenge they consider most important for the coming years*, along with a brief description.

Question 2 Responses. The participants *indicated* 16 *challenges*. The challenges are presented in Table 1; the number in the parentheses that appears in some challenges *indicates the number of participants that mentioned this challenge*. Furthermore, *red font highlights the challenges that are mentioned for the first time in this survey*, i.e., challenges that were not mentioned in the 2020 survey.

The most commonly suggested challenge is the "*Use of LLMs in visualization and analytics*" (voted by 16% of the participants), whereas "*Fairness and Trustworthiness*", as well as "*Visualization for non-expert users*" are the next most common (each voted by 13%). Note that none of the terms LLMs, fairness or trustworthiness is mentioned in the four years ago challenges. Also note that *explainability*, which emerged as one of the most frequently mentioned challenges, is also not mentioned in the 2020 report.

Other common challenges (mentioned by at least two participants) are related to: "*User assistance & guidance*"; "*Understanding what the users need*"; "*High dimensional & stream data*"; "*Progressive data analysis & visualization*"; and "*Immersive visualization*". Among these challenges, "*High dimensional & stream data*", "*Progressive analysis*", and "*Immersive visualization*" appeared for the first time.

4 Emerging Topics

In this section we present the participants' responses regarding the most emerging research topics in Big data visualization and analytics field. The candidate list consists of the topics of interest included in the BigVis call-for-papers.

Table 1: Survey Challenges ⋆

Exploit LLMs$^{(5)}$, Ensure fairness & trustworthiness$^{(4)}$, Enable visualization for non-expert users$^{(4)}$, Offer assistance & guidance$^{(2)}$, Generate explanations$^{(2)}$, Understand what the users need$^{(2)}$, Handle high dimensional & stream data$^{(2)}$, Enable progressive data analysis & visualization$^{(2)}$, Develop immersive visualization systems$^{(2)}$, Provide sustainable insights, Support data abstraction, Implement novel scalable interfaces, Design context-specific visualizations, Formulate fundamental visualization problems, Use surrogate modeling, Develop energy consumption-based solutions

⋆ c$^{(x)}$: x indicates the number of participants that mention the challenge c in the survey. Red font: Challenges mentioned for the first time in the current survey (considering only those indicated by at least two participants).

Question 3

The participants were asked to *select (vote) from a list of candidate topics, up to three topics that they consider the most emerging*.

Question 3 Responses. Figure 3 shows the percentage of the participants' vote for each topic. The topics with the most votes are "*Human–in–the–loop processing*"; "*Interactive & human–centered ML*"; and "*Progressive analytics*", where 31% of the participants select. On the other hand, the topics with the less votes are: "*Scientific visualization*"; "*Setting-oriented visualization*"; and "*Distributed & parallel techniques*", which are voted by 3% of the participants.

5 Discussion

First, the survey *highlights the broad acceptance of the importance of all the challenges identified four years ago* (Question 1), with at least half of the participants strongly agreeing or agreeing on their today's importance.

The results regarding the current challenges (Question 2) revealed the importance of AI–related problems. Notably, *the most frequently mentioned challenges today were entirely absent four years ago*. For example, problems related to LLMs, fairness & trustworthiness, and explanations are some of the newcomers. Furthermore, challenges such as "*Non-expert users*", "*High dimensional & stream data*", "*Progressive analysis*", and "*Immersive visualization*" appeared also for the first time.

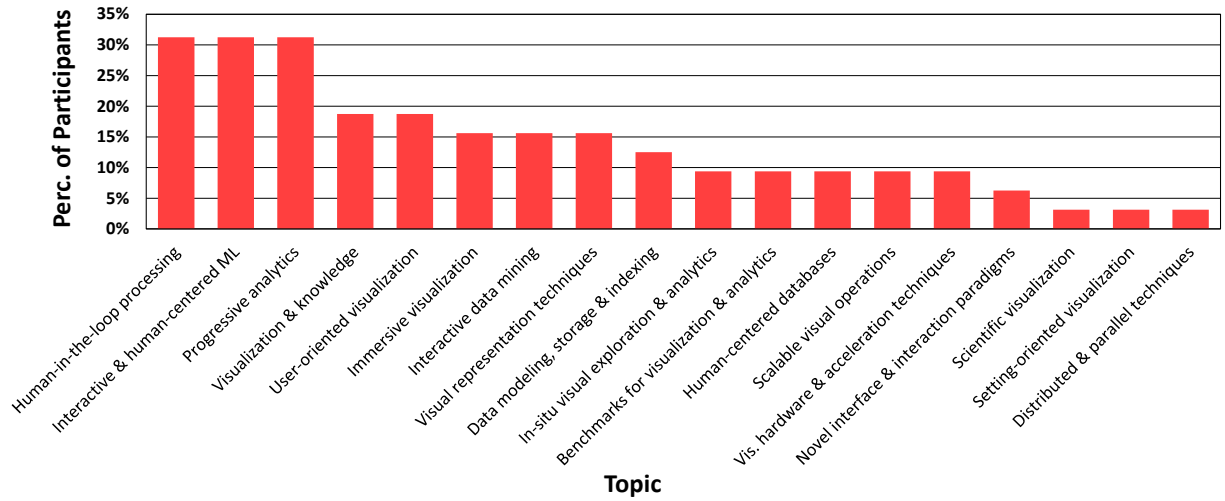


Figure 3: Emerging Topics: The Percentage of Participants that Voted for Each Topic

Further comparison of responses reveals that "*Scalability & efficiency*", the most important challenge in 2020 (Question 1), was not mentioned in Question 2. One possible explanation is that nearly 95% of participants had already rated it in Question 1 as (very) important, reducing the need to highlight it again. Similarly, "*Visual analysis over ML applications*", the second most important challenge in 2020, was absent from Question 2 responses, despite being recognized as one of the most emerging topics (Question 3).

Additional discussion of the current challenges and state-of-the-art approaches can be found in [1, 3–6, 8–15].

References

[1] Sihem Amer-Yahia, Leilani Battle, Yifan Hu, Dominik Moritz, Aditya Parameswaran, Nikos Bikakis, Panos K. Chrysanthis, Guoliang Li, George Papastefanatos, and Lingyun Yu. 2025. Data Exploration and Visual Analytics Challenges in AI Era. ACM SIGMOD Blog, https://wp.sigmod.org/?p=3820.

[2] Gennady L. Andrienko, Natalia V. Andrienko, Steven Mark Drucker, Jean-Daniel Fekete, Danyel Fisher, Stratos Idreos, Tim Kraska, Guoliang Li, Kwan-Liu Ma, Jock D. Mackinlay, Antti Oulasvirta, Tobias Schreck, Heidrun Schumann, Michael Stonebraker, David Auber, Nikos Bikakis, Panos K. Chrysanthis, George Papastefanatos, and Mohamed A. Sharaf. 2020. Big Data Visualization and Analytics: Future Research Challenges and Emerging Applications. In *Workshop on Big Data Visual Exploration & Analytics (BigVis 2020)*.

[3] Natalia V. Andrienko, Gennady L. Andrienko, Linara Adilova, Stefan Wrobel, and Theresa-Marie Rhyne. 2022. Visual Analytics for Human-Centered Machine Learning. *CG&A* (2022).

[4] Rahul C. Basole and Timothy Major. 2024. Generative AI for Visualization: Opportunities and Challenges. *TVCG* (2024).

[5] Leilani Battle and Carlos Scheidegger. 2020. A structured review of data management technology for interactive visualization and analysis. *IEEE TVCG* 27, 2 (2020).

[6] Nikos Bikakis. 2022. Big Data Visualization Tools. In *Encyclopedia of Big Data Technologies, 2nd Ed.* Springer.

[7] Nikos Bikakis, George Papastefanatos, Panos K. Chrysanthis, Olga Papemmanuil, David Auber, Steffen Frey, Issei Fujishiro, Hanna Hauptmann, Shixia Liu, Kwan-Liu Ma, Tobias Schreck, Michael Sedlmair, and Mohamed A. Sharaf. 2024. Visualizing, Exploring and Analyzing Big Data: A 6-Year Story. *ACM SIGMOD Record* 53, 2 (2024).

[8] Muhammad Raees, Inge Meijerink, Ioanna Lykourentzou, Vassilis-Javed Khan, and Konstantinos Papangelis. 2024. From Explainable to Interactive AI: A Literature Review on Current Trends in Human-AI Interaction. *J. Hum. Comput. Stud.* (2024).

[9] Junpeng Wang, Shixia Liu, and Wei Zhang. 2024. Visual Analytics for Machine Learning: A Data Perspective Survey. *IEEE TVCG* 30, 12 (2024).

[10] Qianwen Wang, Zhutian Chen, Yong Wang, and Huamin Qu. 2022. A Survey on ML4VIS: Applying Machine Learning Advances to Data Visualization. *IEEE TVCG* 28, 12 (2022).

[11] Aoyu Wu, Dazhen Deng, Min Chen, Shixia Liu, Daniel A. Keim, Ross Maciejewski, Silvia Miksch, Hendrik Strobelt, Fernanda B. Viégas, and Martin Wattenberg. 2023. Grand Challenges in Visual Analytics Applications. *IEEE CG&A* (2023).

[12] Aoyu Wu, Yun Wang, Xinhuan Shu, Dominik Moritz, Weiwei Cui, Haidong Zhang, Dongmei Zhang, and Huamin Qu. 2022. AI4VIS: Survey on Artificial Intelligence Approaches for Data Visualization. *IEEE TVCG* 28, 12 (2022).

[13] Weikai Yang, Mengchen Liu, Zheng Wang, and Shixia Liu. 2024. Foundation models meet visualizations: Challenges and opportunities. *Comput. Vis. Media* 10, 3 (2024).

[14] Yilin Ye, Jianing Hao, Yihan Hou, Zhan Wang, Shishi Xiao, Yuyu Luo, and Wei Zeng. 2024. Generative AI for visualization: State of the art and future directions. *Vis. Informatics* 8, 1 (2024).

[15] Jun Yuan, Changjian Chen, Weikai Yang, Mengchen Liu, Jiazhi Xia, and Shixia Liu. 2021. A survey of Visual Analytics Techniques for Machine Learning. *Comput. Vis. Media* 7, 1 (2021).

How Diverse Are Our Conference Review Boards?

Sourav S Bhowmick
NTU
Singapore
assourav@ntu.edu.sg

Divesh Srivastava
AT & T Chief Data Office
United States
divesh@research.att.com

ABSTRACT

The diversity of review boards is crucial in improving the overall review quality at academic venues. In this report, we examine the review board characteristics of four major data management conferences across four *diversity dimensions* and over time. Our analysis shows that smaller venues have made significant strides in diversifying their review boards, whereas larger venues encounter difficulties in achieving similar diversity due to overrepresentation in certain dimensions. We emphasize the importance of intensifying efforts to create more diverse and balanced review boards and advocate for the development of tools to support this process.

1. INTRODUCTION

Diversity is widely recognized as a critical factor that plays a significant role for success in many scientific disciplines [13, 15]. It has been reported that groups with diverse members often outperform homogenous groups [13] leading to more impactful work [12, 15]. We expect this to also be true in data management research given its increasing global reach. The data management community, now more than ever, includes people from diverse backgrounds as evident from the geographic diversity of authors in our major venues such as SIGMOD and VLDB. Significant developments have been made possible by individuals from diverse backgrounds as can be seen in the diversity of recipients of major awards in the data management community (*e.g.,* [3–6]) and in general in the computing community (*e.g.,* [1]). For these reasons, the *Diversity, Equity and Inclusion (DEI)* initiative [2] was formed in 2020 to actively promote diversity and inclusion within the data management community [7–9].

Given the benefits of diversity in science, it is desirable for review boards at our major conferences to include experts from varied backgrounds. Such diversity enhances the review process by offering unique perspectives and experiences, while also reducing the risks of groupthink [14] and conformity that are common in homogeneous groups [16]. Thus, a diverse group of experts in a review board has the potential to enhance the overall review quality of a venue, which has a direct impact on the science that emerges from a community. Recently, EDBT 2023 PC chairs reported correlation between the diversity of their review board and the quality of the review process [11].

In this report, we take a concrete step towards analyzing diversity of the review boards of four major data management venues (*i.e.,* SIGMOD, VLDB, EDBT, and ICDE). We focus on the diversity dimensions that can be gleaned from diverse public data sources, *i.e., institutional*, *location*, *country of tertiary/secondary education (COE)*, and *experience*. We conduct a three-year longitudinal study (2023-2025) of the review boards. Note that the formation of the review boards of these venues during this period overlaps with the DEI initiative.

Our study found that while some venues have made progress in promoting diversity on their review boards, this progress is uneven across venues and dimensions. Smaller venues like EDBT have successfully diversified their review boards, while larger venues face challenges in achieving similar diversity. In particular, certain larger venues, such as SIGMOD 2025 and ICDE, show overrepresentation in specific dimensions, failing to reflect the growing diversity of the community or the DEI initiative. We conclude by emphasizing the need for continued efforts to create diverse review boards and the development of data-driven tools to assist program committee (PC) chairs in achieving this goal.

The paper is organized as follows: Section 2 outlines the venues in the study. Section 3 presents the *diversity dimensions* analyzed. Section 4 presents a longitudinal analysis of these dimensions across the venues. Finally, Section 5 concludes with future directions for diverse review board formation.

2. DATASET

The study focuses on the review boards of four

Table 1: The number of reviewers (PC) and meta-reviewers (in brackets) in major data management venues.

Venue	Year		
	2023	2024	2025
SIGMOD	217 [31]	224 [32]	276 [42]
VLDB	216 [35]	206 [38]	314 [50]
EDBT	88 [13]	93 [15]	107 [17]
ICDE	144 [43]	175 [39]	307 [61]

major data management venues–SIGMOD, VLDB, EDBT, and ICDE–from the last three editions (2023-2025). The research has two main objectives: (1) to examine recent trends in the diversity of review boards through a three-year longitudinal study, and (2) to assess the impact of the DEI initiative on the formation of diverse review boards in these venues. The DEI initiative began in 2020 [9], with community awareness in 2021, and the first DEI report was published in June 2022 [9], which coincided with the formation of the review boards for the 2023 conferences. Each of these venues had DEI chairs to promote DEI activities and policies.

The lists of members (meta-reviewers and reviewers (PC)) were received from the PC chairs of respective venues. Table 1 reports the statistics. Since review boards can be dynamic, with members added or removed throughout the review process, the analysis focuses on the aggregated review boards after the final submission cycle for all venues. Observe that review board sizes have generally increased monotonically across all venues over the past three years.

We *manually* retrieved the DBLP addresses of all review board members using Google search. Each review board member for a given venue is uniquely identified by their email address or *DBLP name* which is unique in DBLP[1].

3. CATEGORIES OF DIVERSITY

This article focuses on several dimensions of diversity that can be assessed using publicly available data on review board members. These include *institutional*, *location*, *country of tertiary/secondary education*, and *experience* diversities. Although *gender* and *racial* diversity are important, they are not analyzed due to the lack of publicly available data for many reviewers. Similarly, topics of expertise specified by review board members are not considered, as this information is not publicly accessible.

Institutional diversity focuses on profiling the different institutions that review board members are affiliated with. *Location diversity* refers to the diversity of the countries where review board members are based while serving for a venue. Since the country of origin might not be publicly available, the study uses *country of tertiary/secondary education (COE)* as a proxy, which indicates the country where a reviewer completed their high school or undergraduate education. This information is typically found on a reviewer's homepage or *LinkedIn* page and has been used in review board formation for EDBT 2023 [11].

Lastly, *experience diversity* captures the balance of senior and junior reviewers on a review board. A well-balanced board should have both experienced researchers and junior ones, allowing for expertise while providing opportunities for junior researchers to gain experience. However, there is no universally agreed-upon definition of a "senior" reviewer, as there are no consistent criteria to distinguish between senior and junior reviewers. To assess seniority in terms of research experience, we use *publication age* and *publication venue index* as a proxy for experience.

The *publication age* of a person p in the context of venue v, denoted by $age(p, v)$, is given as follows:

$$age(p, v) = sub\_year(v) - first\_year(p) \quad (1)$$

where $sub\_year(v)$ and $first\_year(p)$ denote the year of the first submission cycle of a venue v and the first year of publication of p, respectively. The larger the value of $age(p, v)$ the more senior p is w.r.t. the number of years of research experience.

The *publication venue index* of a person p w.r.t. venue v, denoted by $venue(p, v)$, is given as follows:

$$venue(p, v) = \sum_{v_i \in V} cnt(p, v_i) \quad (2)$$

where $v \in V$ and $cnt(p, v_i)$ denotes the total number of publications of p in venue v_i. For SIGMOD and VLDB, we choose $V = \{\text{SIGMOD}, \text{VLDB}\}$. For any other venue v (*i.e.,* $v \in \{\text{EDBT}, \text{ICDE}\}$), $V = \{v, \text{SIGMOD}, \text{VLDB}\}$, which includes publications in both the selected venue v and the top venues (SIGMOD, VLDB). The larger the value of $venue(p, v)$ the more publication experience p has w.r.t. v.

Note that both these measures can be computed automatically from DBLP.

4. ANALYSIS

This section analyzes trends related to the dimensions of diversity for reviewers and meta-reviewers at a venue. While meta-reviewers typically do not review submissions, they are included in the study due to their important role in PC formation, often recommending reviewers to the PC chairs. Diverse meta-reviewers can help create a more diverse set of reviewers. In the next section, we shall correlate

[1]In DBLP, homonyms are distinguished from one another by a unique numerical suffix to their name.

Table 2: Institutional diversity.

Venue	Year	IDR (Meta-reviewer)	Top Institution (% of meta-reviewers)	IDR (Reviewer)	Top Institution (% of reviewers)
SIGMOD	2023	0.93	university of waterloo (9.68%)	0.68	national university of singapore (3.23%)
	2024	0.91	microsoft (9.38%)	0.64	microsoft (7.15%)
	2025	0.76	northeastern university (7.14%)	0.63	microsoft (6.88%)
VLDB	2023	0.92	eth zurich, chinese university of hong kong, university of ioannina (5.71%)	0.65	microsoft (3.7%)
	2024	0.92	national university of singapore (5.26%)	0.65	microsoft (4.37%)
	2025	0.94	hong kong university of science and technology (6.0%)	0.61	microsoft (5.25%)
EDBT	2023	0.93	national university of singapore (15.38%)	0.84	microsoft (5.68%)
	2024	1.0	N.A.	0.94	university of calabria, hong kong university of science and technology (guangzhou), university of modena and reggio emilia, universita di bologna, universita degli studi di milano (2.15%)
	2025	0.88	university of waterloo (11.76%), university of ioannina (11.76%)	0.81	athena research center (4.67%)
ICDE	2023	0.88	national university of singapore (6.98%)	0.74	national university of singapore (4.17%)
	2024	0.93	hong kong university of science and technology (10.26%)	0.69	zhejiang university (4.0%)
	2025	0.85	hong kong university of science and technology (8.2%)	0.6	hong kong university of science and technology (3.52%)

Table 3: Location and COE diversity of the review boards.

Venue	Year	Meta-reviewer		Reviewer	
		LDR	CI	LDR	CI
SIGMOD	2023	0.39	0.52	0.11	0.17
	2024	0.31	0.41	0.13	0.15
	2025	0.17	0.38	0.09	0.12
VLDB	2023	0.34	0.37	0.13	0.13
	2024	0.34	0.32	0.11	0.13
	2025	0.32	0.36	0.09	0.11
EDBT	2023	0.69	0.69	0.25	0.22
	2024	0.8	0.53	0.19	0.18
	2025	0.65	0.53	0.17	0.24
ICDE	2023	0.30	0.30	0.15	0.17
	2024	0.33	0.28	0.14	0.15
	2025	0.25	0.28	0.11	0.10

these insights with the recent efforts of the DEI initiative, as outlined in the three annual reports [7–9]. Note that in all figures related to location and COE diversity, the focus is not on clearly displaying individual locations or COEs, but rather on visually emphasizing the skewness of the distributions.

4.1 Institutional Diversity

Since review board members of most venues are affiliated with over 100 institutions, we compute the *institutional diversity ratio (IDR)* rather than using a histogram to analyze institutional diversity. Specifically, it is defined as follows:

$$IDR(v) = \frac{|I_v|}{|R_v|} \tag{3}$$

where R_v is the set of reviewers or meta-reviewers of a venue v and I_v is the set of institutions they are affiliated with. For simplicity, we assume each member $r \in R_v$ is associated with only one institution. Observe that $0 < IDR(v) \leq 1$. A higher $IDR(v)$ value indicates greater institutional diversity in the venue.

Table 2 presents the results of institutional diversity of the four venues. We observe several interesting trends. First, we observe decreasing trend in institutional diversity with time for reviewers (*i.e.,* IDR is decreasing for almost all venues with time). In particular, IDR is lowest for ICDE 2025 (0.6). Similarly, except for VLDB, we generally see decreasing value of IDR for meta-reviewers. An exception is EDBT 2024 which has a perfect IDR value for meta-reviewers (*i.e.,* all meta-reviewers are from distinct institutions) and a high IDR of the PC. Second, we report the institution(s) that most number of (meta)reviewers are associated with in a review board (*i.e., Top Institution* column in Table 2). Interestingly, in ICDE the top institution is dominated by three asian universities. In contrast, the PC of SIGMOD and VLDB is dominated by a non-academic institution (*Microsoft*).

4.2 Location Diversity

Next, we compute the location diversity of the four venues. Since the sizes of the review boards of these venues can vary greatly, we compute the *location diversity ratio (LDR)* as defined as follows:

$$LDR(v) = \frac{|L_v|}{|R_v|} \tag{4}$$

where R_v is the set of reviewers (or meta-reviewers) of a venue v and L_v is the set of locations associated with R_v[2]. Observe that $0 \leq LDR(v) \leq 1$. The higher the value of $LDR(v)$ of a venue the more diverse R_v is w.r.t. locations.

Table 3 presents the LDR values for the venues, ranging from 0.09 to 0.25 for reviewers. EDBT 2023 is the most diverse, while SIGMOD 2025 and VLDB 2025 are the least diverse. All venues, except SIGMOD 2024, show a decline in location diversity over

[2]The individual's location is determined using the institutional domain in their email address; if this fails, the location is manually obtained from their homepage or *LinkedIn* profile.

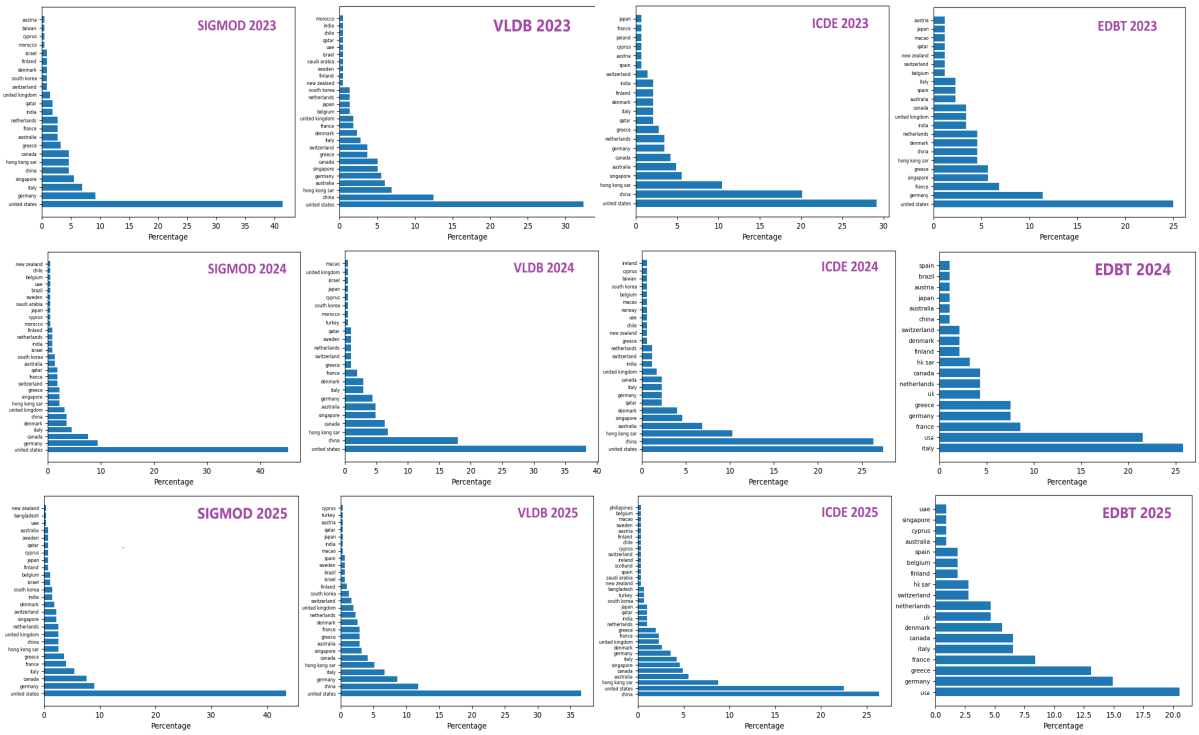


Figure 1: Location distribution of reviewers in SIGMOD, VLDB, ICDE, and EDBT.

Table 4: Top continents of reviewers. The percentage of reviewers is shown in brackets.

Venue	Year		
	2023	2024	2025
SIGMOD	N. America [46.08]	N. America [52.68]	N. America [51.09]
VLDB	N. America [37.5]	N. America [44.66]	N. America [41.36]
EDBT	Europe [44.32]	Europe [66.67]	Europe [66.36]
ICDE	Asia [41.67]	Asia [46.86]	Asia [46.15]

time. For meta-reviewers, LDR values range from 0.17 to 0.8, with SIGMOD showing a decreasing trend, while EDBT has maintained high location diversity over the past three years. In summary, EDBT stands out for having superior location diversity for both reviewers and meta-reviewers over the last three years.

The distributions of location diversity for reviewers are shown in Figure 1. We can observe the following key trends. First, the location of reviewers is primarily dominated by *United States*. For all the three editions of SIGMOD and VLDB 2023, the location distribution is highly skewed, with the difference between the top-2 locations is more than 30%. In contrast, for ICDE and EDBT, the location diversity is more balanced, with the top location varies between 20.6% to 29.2% of the PC. Second, the trend of the top location remains relatively stable for SIGMOD and VLDB, but not for ICDE and EDBT. Notably, EDBT shows a less skewed distribution over time, even as its PC size increases.

Table 4 presents the top continents of the reviewers. The majority of reviewers for SIGMOD and VLDB are from *North America*, while EDBT predominantly features reviewers from *Europe* and ICDE from *Asia*.

Figure 2 shows the distributions of meta-reviewers in representative venues. Similar to the PC distributions, the *United States* is the top location for meta-reviewers in SIGMOD, VLDB, and ICDE (2023 and 2024). However, for EDBT, the top locations vary by year, with the *United States*, *Italy*, and *Germany* leading in different years. The distribution is highly skewed for SIGMOD and VLDB, although VLDB shows increasing diversity in terms of the number of countries. In contrast, ICDE and EDBT have more balanced distributions. There is a downward trend in the skewness of location distribution for VLDB, while SIGMOD shows a steep upward trend. In SIGMOD 2023, the top location comprised 31.25% of meta-reviewers, but this increased to 80.95% in 2025, with the difference between the top two locations rising from 18.75% to 73.81%. As a result, SIGMOD 2025 is the least diverse in terms of meta-reviewer location.

4.3 COE Diversity

Next, we report our observations w.r.t. COE diversity. Note that for any venue, we are not able to find the COE information of at most 0.5% of the re-

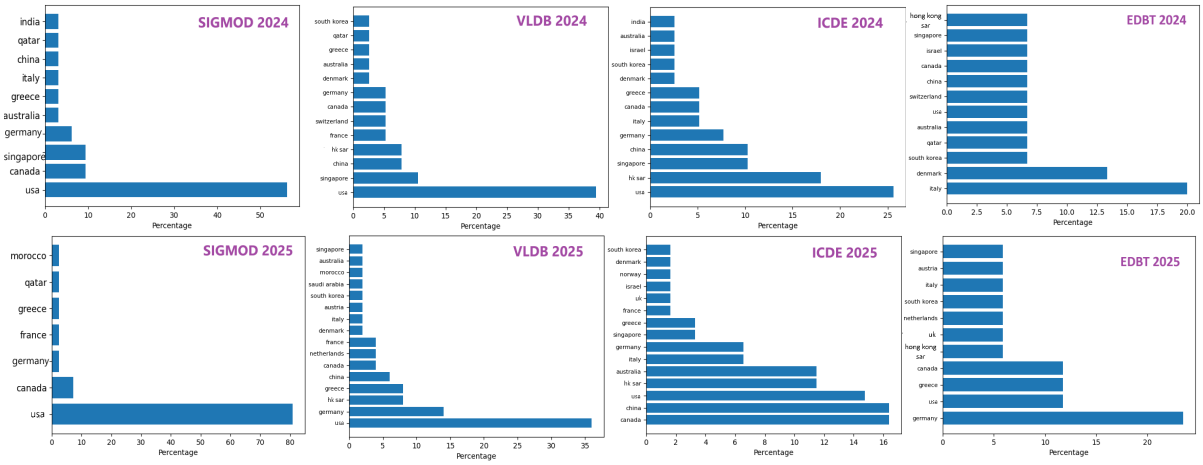


Figure 2: Location distribution of meta-reviewers in SIGMOD, VLDB, ICDE, and EDBT.

viewers. Since other reviewers likely cover the same COEs, their influence on overall trends is negligible.

We compute the *COE Index (CI)* by replacing L_v with C_v in Equation 4, where C_v represents the set of COEs associated with R_v. Table 3 shows that the CI for reviewers ranges from 0.1 to 0.24, with EDBT 2025 being the most diverse and ICDE 2025 the least. Notably, ICDE has shown a declining trend in COE diversity over the past three years. For meta-reviewers, the CI varies from 0.28 to 0.69, with EDBT leading in diversity. Similar to location diversity, SIGMOD exhibits a downward trend in COE diversity over the last three years.

Figure 3 shows the COE diversity distributions for reviewers (PC). We can observe the following key trends. First, except for EDBT, the COE of reviewers in SIGMOD, VLDB, and ICDE is dominated by *China* across all three years. The percentage of such reviewers varies widely, from 22.42% in SIGMOD 2024 to 59.43% in ICDE 2024. Second, COE diversity is significantly skewed for VLDB and ICDE, with the top-2 COEs in VLDB 2024 and ICDE differing in the narrow range of 44.18%–52%, showing a large gap between the top COE and the rest. In contrast, SIGMOD and EDBT have smaller differences, ranging from 8.97% (SIGMOD 2024) to 15.89% (EDBT 2025). Third, while the trend of the top COE remains relatively stable for SIGMOD and EDBT, VLDB and ICDE show an inverted V-shaped trend, with an upward trend from 2023 to 2024, followed by a downward trend in 2025. This trend is more pronounced for VLDB than for ICDE.

Interestingly, the COE distribution of meta-reviewers differs from that of reviewers. Unlike reviewers, no single COE dominates in SIGMOD, VLDB, and EDBT. However, VLDB 2024 and ICDE have similar distribution profiles, with *China* being the top COE across all three years and a significant gap between the top-2 COEs. Additionally, there is a general trend of reduced skewness in the COE distribution for the 2025 edition compared to the previous year.

Table 5: Publication age statistics.

Venue	Year	Reviewer			Meta-reviewer		
		Mean	Med.	S.D.	Mean	Med.	S.D.
SIGMOD	2023	16.95	17	7.53	21	21	5.13
	2024	17.54	17	7.99	23.56	23	7.34
	2025	17.93	17	8.39	22.83	22	7.69
VLDB	2023	14.97	14	7.35	21.71	21	5.69
	2024	15.22	14	7.13	24.24	23.5	7.29
	2025	16.66	16	7.7	21.96	21	7.9
EDBT	2023	15.82	15.5	6.95	22.46	20	9.03
	2024	20.3	20	7.72	24.33	24	7.04
	2025	20.38	20	7.36	27.41	27	6.06
ICDE	2023	14.55	14	7.12	24	23	7.33
	2024	16.62	15	8.51	22.77	22	6.63
	2025	14.25	13	7.71	21.85	20	8.13

4.4 Experience Diversity

Table 5 presents statistics on the publication age of reviewers and meta-reviewers. We observe several interesting trends. First, meta-reviewers consistently have higher publication age than reviewers across all venues, as expected, with the mean and median values closely aligned in most cases. Second, EDBT has a higher average publication age, while VLDB and ICDE have the least experienced review boards w.r.t. publication age over the past three years. Third, except for ICDE, the average publication age of reviewers has been increasing, showing greater experience. However, this trend does not apply to meta-reviewers. All venues, except for EDBT, show an inverted-V trend or decreasing trend in publication age for meta-reviewers (with ICDE showing a decreasing trend). In summary, while review board sizes have grown for all venues, EDBT is the only venue that has consistently seen growth in both board size and publication age.

Table 6 presents the average publication venue index for reviewers across the four venues. The values in brackets indicate the percentage of reviewers with $venue(p, v) = 0$. SIGMOD shows a decreasing

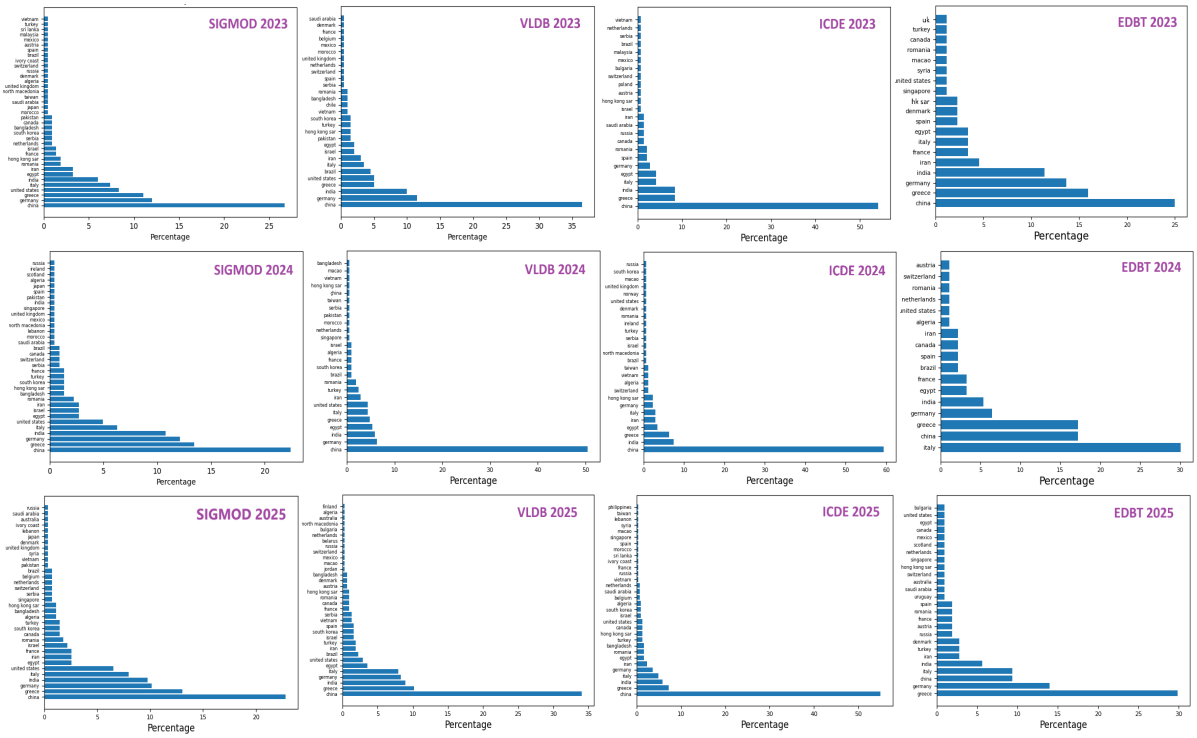


Figure 3: COE distribution of reviewers in SIGMOD, VLDB, ICDE, and EDBT.

trend in the average publication venue index, while the other venues exhibit V-shaped or inverted V-shaped trends. Nevertheless, SIGMOD (resp. ICDE) consistently has higher index compared to VLDB (resp. EDBT). Additionally, the percentage of reviewers with no relevant publications is low for all venues, except for EDBT 2024.

5. REFLECTION & CONCLUSIONS

The DEI initiative in data management initially focused on issues related to participants, authors, and speakers, as well as managing conflicts of interest (COI) for fairer reviewer assignments, as outlined in the 2021 and 2022 reports [8, 9]. The need for diversification of the review boards only surfaced in the most recent report [7]. This article examines the diversity of review boards at four major data management venues. While EDBT has demonstrated superior diversity across the majority of the dimensions of the review boards, larger venues have struggled to maintain this level of consistency. Specifically, there is a noticeable decline in institutional diversity for both reviewers and meta-reviewers across most venues. EDBT shows the most balanced distribution of location and COE diversity among reviewers, while other venues, especially SIGMOD, show significant skewness. Finally, EDBT has the most experienced review boards w.r.t. publication age, while VLDB and ICDE have the least

Table 6: Average publication venue index.

Venue	Year		
	2023	2024	2025
SIGMOD	13.47 [5.07]	13.1 [5.8]	11.93 [4.8]
VLDB	11.14 [4.35]	10.28 [6.09]	11.32 [4.78]
EDBT	11.31 [0]	9.97 [9.68]	14.02 [0.93]
ICDE	14.31 [3.47]	18.75 [2.29]	14.06 [5.21]

experienced, with ICDE showing a declining trend over time. Note that we intentionally avoided speculating on the reasons behind these trends, as the processes involved in forming review boards and the challenges faced are only known to the PC chairs.

The results indicate that review board diversity does not always reflect the growing diversity within the data management community. For instance, while the reviewer database of CLOSET [10] shows over 70 distinct COEs of reviewers in the past five years, less than half of these have been represented on review boards. This highlights the need for more targeted efforts to create diverse and balanced boards. Given the increasing size of review boards and high declination rates for PC invites [11], manually addressing it is difficult for PC chairs. Therefore, the development of *data-driven, PC chair-in-the-loop* tools is essential for efficiently forming diverse and balanced review boards. Lastly, a key goal of a diverse review board is to ensure that submissions are reviewed by diverse experts. Hence, future research will explore the impact of diverse review boards on diversity of reviewer assignments, though access to private assignment data remains a key obstacle.

6. REFERENCES

[1] ACM Awards, https://awards.acm.org/.

[2] Diversity, Equity and Inclusion in Database Conferences, https://dbdni.github.io/.

[3] SIGMOD Awards, https://sigmod.org/sigmod-awards/.

[4] VLDB Test of Time Award. https://vldb.org/awards_10year.html.

[5] VLDB Early Career Award. https://vldb.org/awards_early_career.html.

[6] Women in Database Research Award. https://vldb.org/awards_women_in_DB.html.

[7] S. Amer-Yahia, D. Agrawal, Y. Amsterdamer, S. S. Bhowmick, R. Borovica-Gajic, J. Camacho-Rodríguez, J. Cao, B. Catania, P. K. Chrysanthis, C. Curino, A. E. Abbadi, A. Floratou, J. Freire, S. Idreos, V. Kalogeraki, S. Maiyya, A. Meliou, M. Mohanty, F. Özcan, L. Peterfreund, S. Sahri, S. Sellami, R. Shraga, U. Sirin, W.-C. Tan, B. Thuraisingham, Y. Tian, G. Vargas-Solar, M. Zhang, W. Zhang. Diversity, Equity and Inclusion Activities in Database Conferences: A 2023 Report. *SIGMOD Record*, 53(2), June 2024.

[8] S. Amer-Yahia, D. Agrawal, Y. Amsterdamer, S. S. Bhowmick, A. Bonifati, R. Borovica-Gajic, J. Camacho-Rodríguez, B. Catania, P. K. Chrysanthis, C. Curino, J. Darmont, G. Dobbie, A. E. Abbadi, A. Floratou, J. Freire, A. Jindal, V. Kalogeraki, S. Maiyya, A. Meliou, M. Mohanty, B. Omidvar-Tehrani, F. Özcan, L. Peterfreund, W. Rahayu, S. Sadiq, S. Sellami, U. Sirin, W.-C. Tan, B. Thuraisingham, Y. Tian, P. Tözün, G. Vargas-Solar, N. J. Yadwadkar, V. Zakhary, M. Zhang. Diversity, Equity and Inclusion Activities in Database Conferences: A 2022 Report. *SIGMOD Rec.* 52(2): 38-42, June 2023.

[9] S. Amer-Yahia, Y. Amsterdamer, S. S. Bhowmick, A. Bonifati, P. Bonnet, R. Borovica-Gajic, B. Catania, T. Cerquitelli, S. Chiusano, P. K. Chrysanthis, C. Curino, J. Darmont, A. E. Abbadi, A. Floratou, J. Freire, A. Jindal, V. Kalogeraki, G. Koutrika, A. Kumar, S. Maiyya, A. Meliou, M. Mohanty, F. Naumann, N. S. Noack, F. Özcan, L. Peterfreund, W. Rahayu, W.-C. Tan, Y. Tian, P. Tözün, G. Vargas-Solar, N. J. Yadwadkar, M. Zhang. Diversity and Inclusion Activities in Database Conferences: A 2021 Report. *SIGMOD Rec.* 51(2): 69-73, June 2022.

[10] S. S. Bhowmick. CLOSET: Data-Driven COI Detection and Management in Peer-Review Venues. *Commun. ACM*, 66(7): 70-71, July 2023.

[11] S. S. Bhowmick, K. Hose. Data-driven PC-chair-in-the-loop Formation of Program Committees: An EDBT 2023 Experience. *SIGMOD Rec.*, 53(2), 68-74, June 2024.

[12] R. B. Freeman, W. Huang. Collaborating with People Like me: Ethnic Coauthorship Within the United States. *J. Labor Econ.*, 33, S1, July 2015.

[13] L. Hong, S. E. Page. Groups of Diverse Problem Solvers can Outperform Groups of High-ability Problem Solvers. *In Proc. of the National Academy of Sciences*, 101, 46, Nov. 2004.

[14] W. H. Whyte, Jr. (March 1952). Groupthink. *Fortune.* pp. 114–117, 142, 146.

[15] Y. Yang *et al.*. Gender-diverse Teams Produce more Novel and Higher-impact Scientific Ideas. *In Proc. of the National Academy of Sciences*, 119, 36, Sept. 2022.

[16] P. Zimbardo, R. Johnson, V. McCann. Psychology Core Concepts. *Pearson Education, Inc.*, 8th Edition, 2016.