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, Rachel Pottinger, Sayan Ranu, Cyrus Shahabi, 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)
Jeffrey Ullman (1996)
Raghu Ramakrishnan (1999)
Daniel Rosenkrantz (2001)
Surajit Chaudhuri (2004)
Hans-Jörg Schek (2007)
David Lomet (2010)
H.V. Jagadish (2013)
Samuel Madden (2016)
Ahmed Elmagarmid (2019)
Stratos Idreos (2020)
Dennis Shasha (2020)
K. Selcuk Candan (2023)
Gio Wiederhold (1995)
Avi Silberschatz (1997)
Michael Carey (2000)
Richard Snodgrass (2002)
Hongjun Lu (2005)
Klaus R. Dittrich (2008)
Gerhard Weikum (2011)
Kyu-Young Whang (2014)
Yannis E. Ioannidis (2017)
Philipe Bonnet (2020)
Stefan Manegold (2020)
Divesh Srivastava (2021)
Sihem Amer-Yahia (2024)
Yahiko Kambayashi (1995)
Won Kim (1998)
Laura Haas (2000)
Michael Ley (2003)
Tamer Özsu (2006)
Beng Chin Ooi (2009)
Marianne Winslett (2012)
Curtis Dyreson (2015)
Z. Meral Özsoyoğlu (2018)
Juliana Freire (2020)
Ioana Manolescu (2020)
Christian S. Jensen (2022)
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
- **2025** *Winner:* Peng Li, *Honorable Mentions:* Xuanhue Zhou, Aecio Santos, and Meghdad Kurmanji

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 September 2025 issue of the ACM SIGMOD Record!

This issue starts with the Database Principles column featuring an article by Bienvenu and colleagues on recent advances in logic-based entity resolution. Declarative logic-based methods have been used to capture and exploit relational dependencies, which makes them well suited for handling complex multirelational setting arising in collective entity resolution. The article focuses on logic-based approaches to entity resolution that are collective, declarative, and justifiable. The authors present recent foundational and conceptual advances in this space, discuss the Lace framework as a prominent example, and outline directions for future work in logic-based collective entity resolution.

The Reminiscences on Influential Papers column, edited by Pınar Tözün, presents contributions by Viktor Leis, Anja Gruenheid, Paris Carbone, and Eleni Tzirita Zacharatou.

The Advice to Mid-Career Researchers column presents a contribution by Christian S. Jensen, who shares his thoughts and experiences on research as a social activity, on engaging in community efforts, on providing service to the scientific community, and on seeking flow. The article provides advice on many aspects of the mid-career stage of life, including balancing continuity and renewal, searching for unexplored territories while maintaining continuity of ideas, the need for and approaches to sometimes saying no - and on the benefits of always being nice!

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 Zsolt István on the performance-security trade-off in analytics on shared data. The article considers approaches based on secure multi-party computation, which offer impressive security protections in decentralized settings, albeit in potentially inefficient ways. In this context, the author proposes to approach security-efficiency tradeoffs at the operator and query level, and presents a case study that focuses on protecting the sizes of intermediate results passed between query operators.

The Distinguished Profiles column features an interview with Sihem Amer-Yahia, a Silver Medal Research Director at the French National Center for Scientific Research (CNRS) and Deputy Director of the Laboratoire d'Informatique de Grenoble, one of the largest research labs in Computer Science in France, with CNRS and INRIA Researchers and University Professors. She has won many awards, including the 2024 IEEE TCDE Impact Award, the ACM SIGMOD Contributions Award, and the VLDB Women in Database Research Award. In the interview, Sihem discusses the research contributions that she is most proud of, shares her thoughts on XML and DB/IR research, and outlines the work that she feels needs to be done by the community on involving individuals and groups as first-class citizens. She also talks about the DEI initiative, its history and evolution, and her perspective on her involvement in the movement since its inception. In the context of her life experience, she provides advice to women entering DB research, and also talks about her having lived and worked in different countries, about the sense of purpose, and about dancing. The interview is closed by Sihem discussing her perspective on the future of data-management research and educational practices.

The issue closes with the Reports column, which presents an article by Barret and colleagues on the diversity, equity and inclusion (DEI) activities in database conferences in 2024. The article articulates the goal of the community with respect to these activities, presents the new structure of the DEI initiative, and shares the DEI statistics for 2024 and the results of a 2024 SIGMOD survey on this topic.

The authors also discuss the 2024 progress in conference COI management, reflect on the overall 2024 progress, and present DEI plans for 2025.

On behalf of the SIGMOD Record Editorial board, I hope that you enjoy reading the September 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

September 2025

Past SIGMOD Record Editors:

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

Recent Advances in Logic-Based Entity Resolution

Meghyn Bienvenu
Univ. Bordeaux, CNRS, LaBRI

Gianluca Cima
Sapienza University of Rome

Víctor Gutiérrez-Basulto
Cardiff University

Yazmín Ibáñez-García
Cardiff University

Zhiliang Xiang
Cardiff University

ABSTRACT

Entity resolution (ER) is a central task in data quality, which is concerned with identifying pairs of distinct constants or tuples that refer to the same real-world entity. Declarative approaches, based upon logical rules and constraints, are a natural choice for tackling complex, collective ER tasks involving the joint resolution of multiple entity types across multiple tables. This paper provides an overview of recent advances in logic-based entity resolution, with a particular focus on the LACE framework, first introduced at PODS'22 and subsequently extended with additional features (IJCAI'23, KR'23) and equipped with an answer set programming-based implementation (KR'24, KR'25).

1 Introduction

Entity resolution (ER) is a key data quality task that seeks to identify distinct constants that refer to the same real-world entity [54]. A wide range of ER approaches have been proposed, differing in their assumptions, the nature of the data they handle, and the techniques they employ [21]. In the context of relational databases, ER has traditionally focused on matching records based on field-level similarity [47], which is why it is also known as record linkage [35]. For example, in a bibliographic database, *two author records might be matched if their email addresses are similar*. A more expressive and general approach, known as collective ER [9, 24], performs joint resolution of entity references or values of multiple types across multiple tables. For instance, *merging two author entities may lead to the inference that their associated paper IDs should also be merged.* Most existing approaches to ER focus on single-pass matching of tuples within a single table or between a pair of tables, and machine learning methods have obtained remarkable results [41] for such settings. On the other hand, declarative methods based on logical rules and constraints are able to capture and exploit relational dependencies, making them well-suited for handling complex multi-relational settings arising in collective ER.

In this paper, we examine declarative approaches to collective entity resolution[1], with a particular focus on the LACE framework. We designed LACE to satisfy three natural desiderata: being *collective*, *declarative*, and *justifiable*. Specifically, our approach (i) supports complex interdependencies between merges of different entities, (ii) adopts a declarative language based on logical rules and constraints, and (iii) provides clear justifications for why two constants are considered to represent the same entity. While the collective and declarative aspects have received considerable attention in the literature, the notion of justifiability remains relatively underexplored, despite being a crucial step toward building more advanced explanation capabilities and, ultimately, more responsible technologies [42].

As a declarative language, LACE shares several design principles with existing logic-based frameworks. Inspired by the Dedupalog framework [2], it employs both hard and soft rules to specify conditions under which pairs of constants must or may be merged. For instance, statements such as *every paper has a single corresponding author* and *conferences with similar names are likely to be the same* can be naturally expressed using hard and soft rules, respectively. Beyond rules, LACE specifications can also include denial constraints [7], which ensure the consistency of the resulting instance by restricting inadmissible combinations of merges. For example, one may enforce that *an author's name can appear only once in the list of authors of a paper.* In line with entity resolution approaches based on matching dependencies (MDs) [8, 27, 32] and their extensions, such as relational MDs [4, 5] or entity-enhancing rules (REEs) [24, 33], LACE adopts a dynamic semantics in which rule bodies are evaluated over the evolving instance, i.e., the instance resulting from merges that have already been derived.

[1]For a more extensive discussion of ER methods, the interested reader is referred to [21].

The dynamic semantics is key to obtaining a collective yet justifiable framework: merges can trigger additional merges, potentially in a recursive manner, while still guaranteeing that all merges occurring in a solution are *justified*, in the sense that it is possible to trace back how each merge was obtained via a sequence of rule applications.

Another important design consideration concerns the nature of the constants being merged. When constants represent entity references (e.g., author names or paper IDs), a global semantics, where all occurrences of matched constants are merged (not just those directly involved in deriving the match) is particularly well suited. In contrast, when dealing with data values, a local merging semantics, one that considers the context in which a value occurs, is often more appropriate. For example, some occurrences of 'J. Smith' may refer to 'Joe Smith', while others may correspond to 'Jane Smith'; merging all instances globally in such cases would lead to incorrect resolutions. Various efforts have already been dedicated to studying both approaches. For instance, MDs provide a principled logical formalism for merging values [8, 27, 32], adopting a local semantics. On the other hand, Dedupalog, MRLs, and the declarative framework for *entity linking* [17] (henceforth referred to as EL) rely on a global semantics. Despite substantial work on each of these approaches, most existing frameworks focus on one or the other. This was also the case for the Lace framework, which initially only supported global merges of entity references [11] but was subsequently extended [13] to also handle local merges of data values. Note that, contemporaneously to Lace, the CERQ framework supporting both local and global merges was independently developed [26].

Another distinction among logic-based approaches to ER lies in the nature of their solutions or outputs. A key aspect in this regard is whether the approach produces a single solution or a set of alternative solutions.[2] As in the EL approach [17], we consider not just one solution but a space of preferred solutions. In Lace, this space arises naturally from the role of denial constraints, which restrict which merges can co-occur, thereby introducing meaningful choices. Also in line with EL, we can naturally define the notions of *certain* and *possible* merges, referring respectively to those merges that appear in all, or in some, of the preferred solutions.

[2]Here, a solution can be roughly viewed as a set of constant pairs that are judged to refer to the same entity. Naturally, outside logic-based approaches, solutions may take other forms, for example, expressing the likelihood that two constants refer to the same entity.

We argue that the successful adoption of any ER framework depends on the availability of an accompanying implementation. To this end, we have developed the answer set programming (ASP)-based systems ASPen and ASPen$^+$, grounded in the foundational result that Lace solutions can be faithfully encoded as ASP stable models. ASPen$^+$ supports the full range of Lace features and further extends the framework by exploring various optimality criteria; not only prioritizing solutions that maximize the number of merges (w.r.t. set inclusion), but also enabling other natural forms of preference.

Organization After reviewing the necessary background in Section 2, we introduce in Section 3 the fundamentals of Lace, including its syntax, semantics, properties, and alternative optimality criteria. In Section 4, we define the central decision problems and analyse their computational complexity. Section 5 discusses the practical implementation of Lace using the answer set programming systems ASPen and ASPen$^+$. In Section 6, we present Replace, a holistic framework that integrates ER and repairing. Section 7 overviews the broader landscape of logic-based ER approaches. Finally, Section 8 offers perspectives for future work.

2 Preliminaries

Databases We assume that *constants* are drawn from three infinite and pairwise disjoint sets: a set O of *object constants* (or *objects*), serving as references to real-world entities (e.g. paper and author ids), a set V of *value constants* (or *values*) from the considered datatypes (e.g. strings for names of authors and paper titles, dates for time of publication), and a set TID of *tuple identifiers (tids)*.

A *(database) schema* $\mathcal{S}$ consists of a finite set of *relation symbols*, each having an associated arity $k \in \mathbb{N}$ and type vector $\{\mathsf{O}, \mathsf{V}\}^k$. We use $R/k \in \mathcal{S}$ to indicate that the relation symbol R from $\mathcal{S}$ has arity k, and denote by $\mathbf{type}(R, i)$ the ith element of R's type vector. If $\mathbf{type}(R, i) = \mathsf{O}$ (resp. V), we call i an *object (resp. value) position* of R.

An *$\mathcal{S}$-database* is a finite set D of *facts* of the form $R(t, c_1, \ldots, c_k)$, where $R/k \in \mathcal{S}$, $t \in \mathsf{TID}$, and $c_i \in \mathbf{type}(R, i)$ for $1 \leq i \leq k$. We require that each $t \in \mathsf{TID}$ occurs in at most one fact of D. Abusing notation, we will sometimes use t to refer the unique fact with tid t and use $t[j]$ for the constant in the jth position of t (tid arguments occupy position 0, and 'regular' arguments of R/k are in positions $1, \ldots, k$). The set of constants (resp. objects) occurring in D is denoted $\mathsf{Dom}(D)$ (resp. $\mathsf{Obj}(D)$), and the set $\mathsf{Cells}(D)$ of *(value) cells* of D is defined as $\{\langle t, i\rangle \mid t \in D, t[I] \in \mathsf{V}\}$.

Queries We consider *conjunctive queries (CQs)* of the form $q(\mathbf{x}) = \exists \mathbf{y}.\varphi(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ and $\mathbf{y}$ are disjoint tuples of variables, and $\varphi(\mathbf{x}, \mathbf{y})$ is a conjunction of relational atoms $R(u_0, u_1, \ldots, u_k)$, where $R/k \in \mathcal{S}$, $u_0 \in \mathsf{TID} \cup \mathbf{x} \cup \mathbf{y}$, and for every $1 \leq i \leq k$: $u_i \in \mathsf{O} \cup \mathsf{V} \cup \mathbf{x} \cup \mathbf{y}$ and $u_i \in \mathsf{O} \cup \mathsf{V}$ implies $u_i \in \mathbf{type}(R, i)$. When formulating entity resolution rules and constraints, we shall also consider extended forms of CQs that may contain inequality atoms or atoms built from a set of binary *similarity relations*. Note that such atoms will not contain the tid position and have a fixed meaning[3]. Moreover, we impose a standard safety condition: each variable occurring in an inequality or similarity atom must also occur in some relational atom (in a value position, in the case of similarity atoms). As usual, the *arity* of $q(\mathbf{x})$ is the length of $\mathbf{x}$, and queries of arity 0 are called *Boolean*. Given an n-ary query $q(x_1, \ldots, x_n)$ and n-tuple of constants $\mathbf{c} = (c_1, \ldots, c_n)$, we denote by $q[\mathbf{c}]$ the Boolean query obtained by replacing each x_i by c_i. We denote by $\mathsf{vars}(q)$ (resp. $\mathsf{cons}(q)$) the set of variables (resp. constants) in q. and will use set notation for queries when convenient.

Constraints Our framework will also employ *denial constraints* (DCs) [7, 28] which take the form $\exists \mathbf{y}.\varphi(\mathbf{y}) \rightarrow \bot$, where $\exists \mathbf{y}.\varphi(\mathbf{y})$ is a Boolean CQ with inequalities, whose relational atoms use relation symbols from the considered schema $\mathcal{S}$. DCs notably generalize the well-known class of *functional dependencies (FDs)*. To simplify the presentation, we sometimes omit the quantifiers from DCs.

3 LACE Framework

In this section, we present LACE[4], a Logical Approach to Collective Entity resolution, designed to satisfy the desiderata laid out in Section 1.

3.1 Syntax of LACE

In LACE, rules are used to describe conditions under which two constants must or may be identified (we use the term 'merge' to speak of identified pairs). These come in two flavours, depending on whether the considered constants are objects or values.

Rules for Objects To formalize the resolution of object pairs (i.e., references to real-world entities) that denote the same underlying entity, we employ *hard and soft rules for objects* (over a schema $\mathcal{S}$), which take respectively the following forms:

$$q(x, y) \Rightarrow \mathsf{EqO}(x, y) \quad q(x, y) \dashrightarrow \mathsf{EqO}(x, y)$$

where $q(x, y)$ is a CQ whose atoms may use relation symbols from $\mathcal{S}$ as well as similarity relations and whose free variables x and y occur only in object positions. Intuitively, the above hard (resp. soft) rule states that (o_1, o_2) being an answer to q provides sufficient (resp. reasonable) evidence for concluding that o_1 and o_2 refer to the same real-world entity. Note that rules for objects use a special relation symbol EqO (not in schema $\mathcal{S}$) in the rule head.

Rules for Values To formalize *local identifications* between distinct representations of the same information, we introduce *hard and soft rules for values*, which take respectively the following forms:

$$q(x_t, y_t) \Rightarrow \mathsf{EqV}(\langle x_t, i\rangle, \langle y_t, j\rangle)$$
$$q(x_t, y_t) \dashrightarrow \mathsf{EqV}(\langle x_t, i\rangle, \langle y_t, j\rangle)$$

where $q(x_t, y_t)$ is a CQ whose atoms may use relation symbols from the considered schema $\mathcal{S}$ as well as similarity relations, variables x_t and y_t each occur once in q in position 0 of (not necessarily distinct) relational atoms with relations $R_x \in \mathcal{S}$ and $R_y \in \mathcal{S}$, respectively, and i and j are value positions of R_x and R_y, respectively. Intuitively, such a hard (resp. soft) rule states that a pair of tids (t_1, t_2) being an answer to q provides sufficient (resp. reasonable) evidence for concluding that the values in cells $\langle x_t, i\rangle$ and $\langle y_t, j\rangle$ are non-identical representations of the same information. Rules for values use head relation EqV (not in $\mathcal{S}$ and distinct from EqO).

To refer to a generic (hard or soft) rule, we use the arrow symbol $\rightarrow$ (which can stand for $\Rightarrow$ or $\dashrightarrow$). For the sake of brevity, we usually omit existential quantifiers of variables in rule bodies.

ER Specifications In addition to rules for indicating mandatory or likely merges, LACE specifications may also include denial constraints, which serve to define what counts as a legal (or consistent) database and can help to block incorrect merges.

DEFINITION 1. *A LACE entity resolution (ER) specification Σ for schema $\mathcal{S}$ takes the form $\Sigma = \langle \Gamma_O, \Gamma_V, \Delta\rangle$, where $\Gamma_O = \Gamma_h^o \cup \Gamma_s^o$ is a finite set of hard and soft rules for objects, $\Gamma_V = \Gamma_h^v \cup \Gamma_s^v$ is a finite set of hard and soft rules for values, and Δ is a finite set of denial constraints, all over $\mathcal{S}$.*

EXAMPLE 1. *The schema $\mathcal{S}_{\mathsf{ex}}$, database D_{ex}, and ER specification $\Sigma_{\mathsf{ex}} = \langle \Gamma_{\mathsf{ex}}^O, \Gamma_{\mathsf{ex}}^V, \Delta_{\mathsf{ex}}\rangle$ of our running example are given in Figure 1. Informally, the*

[3]Similarity relations are typically defined by applying a similarity metric, e.g. edit distance, and keeping those pairs of values whose score exceeds a given threshold.

[4]We present here the version of LACE from [13], which extends the original LACE framework [11] to support local merges of values, rather than only global merges of objects, as was the case in [11]. Note that this version of the framework is referred to as LACE$^+$ in [13].

denial constraint δ_1 is an FD saying that an author id is associated with at most one author name, while the constraint δ_2 forbids the existence of a paper written by the chair of the conference in which the paper was published. The hard rule ρ_1^o states that if two author ids have the same name and the same institution, then they refer to the same author. The soft rule σ_1^o states that authors who wrote a paper in common and have similar names are likely to be the same. Finally, the hard rule ρ_1^v locally merges similar names associated with the same author id.

3.2 Semantics of LACE Specifications

As the aim is to identify pairs of objects (resp. occurrences of values) that denote the same real-world entity (resp. represent the same value), we will be interested in solutions taking the form of a *pair of equivalence relations* $\langle E, V\rangle$, over the sets $\mathsf{Obj}(D)$ and $\mathsf{Cells}(D)$ respectively, giving the merged pairs of objects and value cells. To satisfy our desiderata, we must ensure that the set of merges present in a solution can be justified by appealing to the rules and is coherent w.r.t. the expressed constraints. The idea will thus be to define solutions in terms of sequences of rule applications that lead to a database satisfying all hard rules and denial constraints. Importantly, rule bodies and constraints will be evaluated with respect to the database induced by the current pair of equivalence relations, in order to exploit the previously derived object and cell merges.

To make this formal, we need to clarify the notion of 'induced database' obtained by modifying the initial database to 'implement' a given set of merges. We might be tempted to simply pick a representative from each equivalence class and replace every constant with its representative. While such an approach can be used to treat global merges of (as was done in [11]), it cannot account for the local nature of cell-level merges of values. For this reason, we shall work with an extended form of database, where the arguments are *sets of constants*.

Definition 2. *Given an $\mathcal{S}$-database D, equivalence relation E over $\mathsf{Obj}(D)$, and equivalence relation V over $\mathsf{Cells}(D)$, we denote by $D_{E,V}$ the* (extended) database induced by D, E, and V, *which is obtained from D by replacing:*

- *each tid t with the singleton set $\{t\}$,*
- *each occurrence of $o \in \mathsf{Obj}(D)$ by $\{o' \mid (o, o') \in E\}$,*
- *each value in a cell $\langle t, i\rangle \in \mathsf{Cells}(D)$ with the set of values $\{t'[i'] \mid (\langle t, i\rangle, \langle t', i'\rangle) \in V\}$.*

It remains to specify how queries in rule bodies and constraints are to be evaluated over such induced databases. First, we need to say how similarity predicates are extended to sets of constants. We propose that $C_1 \approx C_2$ is satisfied whenever there are $c_1 \in C_1$ and $c_2 \in C_2$ such that $c_1 \approx c_2$, since the elements of a set provide different possible representations of a value. Second, we must take care when handling join variables in value positions. Requiring all occurrences of a variable to map to the same set is too strong, e.g. it forbids us from matching {J. Smith, Joe Smith} with {J. Smith}. We require instead that the intersection of all sets of constants assigned to a given variable is non-empty.

Definition 3. *A Boolean query q (possibly containing similarity and inequality atoms) is* satisfied *in $D_{E,V}$, denoted $D_{E,V} \models q$, if there exists a function $h : \mathsf{vars}(q) \cup \mathsf{cons}(q) \to 2^{\mathsf{Dom}(D)} \setminus \{\emptyset\}$ and functions $g_\pi : \{0, \ldots, k\} \to 2^{\mathsf{Dom}(D)}$ for each k-ary relational atom $\pi \in q$, such that:*

1. *for every $a \in \mathsf{cons}(q)$, $h(a) = \{a\}$, and for every $z \in \mathsf{vars}(q)$, $h(z)$ is the intersection of all sets $g_\pi(i)$ such that z is the ith argument of π;*
2. *for every relational atom $\pi = R(u_0, u_1, \ldots, u_k) \in q$, $R(g_\pi(0), g_\pi(1), \ldots, g_\pi(k)) \in D_{E,V}$, and for every $1 \leq i \leq k$, if $u_i \in \mathsf{cons}(q)$, then $u_i \in g_\pi(i)$;*
3. *for every atom $z \neq z' \in q$: $h(z) \cap h(z') = \emptyset$;*
4. *for every atom $u \approx u' \in q$: there exist $c \in h(u)$ and $c' \in h(u')$ such that $c \approx c'$.*

For non-Boolean queries, the set $q(D_{E,V})$ of answers to $q(\mathbf{x})$ contains tuples $\mathbf{c}$ s.t. $D_{E,V} \models q[\mathbf{c}]$.

Observe that the functions g_π make it possible to map different occurrences of the same variable z to different sets of constants, with Point 1 ensuring these sets have a non-empty intersection, $h(z)$. It is this intersected set, storing the common values for z, that is used to evaluate inequality and similarity atoms. Note that when a constant c occurs in a relational atom, the set assigned to the position where c occurs must contain c.

The preceding definition of satisfaction of queries straightforwardly extends to constraints and rules:

- $D_{E,V} \models \exists \mathbf{y}.\varphi(\mathbf{y}) \to \bot$ iff $D_{E,V} \not\models \exists \mathbf{y}.\varphi(\mathbf{y})$
- $D_{E,V} \models q(x, y) \to \mathsf{EqO}(x, y)$ iff $q(D_{E,V}) \subseteq E$
- $D_{E,V} \models q(x_t, y_t) \to \mathsf{EqV}(\langle x_t, i\rangle, \langle y_t, j\rangle)$ iff $(t_1, t_2) \in q(D_{E,V})$ implies $(\langle t_1, i\rangle, \langle t_2, j\rangle) \in V$,

where symbol $\to$ may be either $\Rightarrow$ or $\dashrightarrow$. We write $D_{E,V} \models \Lambda$ iff $D_{E,V} \models \lambda$ for every $\lambda \in \Lambda$.

Author(tid, aid, name, inst)

tid	aid	name	inst
t_1	a_1	J. Smith	Sapienza
t_2	a_2	Joe Smith	Oxford
t_3	a_3	J. Smith	NYU
t_4	a_4	Joe Smith	NYU
t_5	a_5	Joe Smith	Sapienza
t_6	a_6	Min Lee	CNRS
t_7	a_7	M. Lee	UTokyo
t_8	a_8	Myriam Lee	Cardiff

Wrote(tid, aid, pid)

tid	aid	pid
t_{14}	a_1	p_1
t_{15}	a_2	p_1
t_{16}	a_3	p_2
t_{17}	a_6	p_3
t_{18}	a_7	p_3
t_{19}	a_7	p_4
t_{20}	a_8	p_4
t_{21}	a_6	p_5

Paper(tid, pid, title, conf, chr)

tid	pid	title	conf	chr
t_9	p_1	Logical Framework for ER	PODS'21	a_6
t_{10}	p_2	Rule-based approach to ER	ICDE'19	a_4
t_{11}	p_3	Query Answering over DLs	KR'22	a_1
t_{12}	p_4	CQA over DL Ontologies	IJCAI'21	a_1
t_{13}	p_5	Semantic Data Integration	AAAI'22	a_8

$\delta_1 = \text{Author}(t,a,n,i) \wedge \text{Author}(t',a,n',i') \wedge n \neq n' \rightarrow \bot$

$\delta_2 = \text{Paper}(t,p,ti,c,a) \wedge \text{Wrote}(t',a,p) \rightarrow \bot$

$\rho_1^o = \text{Author}(t,x,n,i) \wedge \text{Author}(t',y,n,i) \Rightarrow \mathsf{EqO}(x,y)$

$\rho_1^v = \text{Author}(t,a,n,i) \wedge \text{Author}(t',a,n',i') \wedge n \approx n' \Rightarrow \mathsf{EqV}(\langle t,2\rangle, \langle t',2\rangle)$

$\sigma_1^o = \text{Author}(t,x,n,i) \wedge \text{Author}(t',y,n',i') \wedge n \approx n' \wedge \text{Wrote}(t'',x,p) \wedge \text{Wrote}(t''',y,p) \dashrightarrow \mathsf{EqO}(x,y)$

Figure 1: Schema $\mathcal{S}_{\mathsf{ex}}$, database D_{ex}, and specification $\Sigma_{\mathsf{ex}} = \langle \Gamma^O_{\mathsf{ex}}, \Gamma^V_{\mathsf{ex}}, \Delta_{\mathsf{ex}} \rangle$ with $\Gamma^O_{\mathsf{ex}} = \{\rho_1^o, \sigma_1^o\}$, $\Gamma^V_{\mathsf{ex}} = \{\rho_1^v\}$, and $\Delta_{\mathsf{ex}} = \{\delta_1, \delta_2\}$. Similarity relation $\approx$ is defined so names of authors a_1, a_2, a_3, a_4, and a_5 are all pairwise similar, and the names of a_6 and a_8 are both similar to the name of a_7 (but not similar to each other).

With these notions in hand, we can formally define solutions to an ER specification and dataset. The definition employs the notation $\mathsf{EqRel}(P, S)$, giving the least equivalence relation on set S that contains all pairs in P (i.e. minimally extending P to satisfy reflexivity, symmetry, and transitivity).

Definition 4. *Given an ER specification $\Sigma = \langle \Gamma_O, \Gamma_V, \Delta \rangle$ over schema $\mathcal{S}$ and an $\mathcal{S}$-database D, we call $\langle E, V \rangle$ a* candidate solution for (D, Σ) *if it satisfies one of the following three conditions:*

- $E = \mathsf{EqRel}(\emptyset, \mathsf{Obj}(D))$ *and* $V = \mathsf{EqRel}(\emptyset, \mathsf{Cells}(D))$
- $E = \mathsf{EqRel}(E' \cup \{(o, o')\}, \mathsf{Obj}(D))$, *where $\langle E', V \rangle$ is a candidate solution for (D, Σ) and $(o, o') \in q(D_{E,V})$ for some $q(x,y) \rightarrow \mathsf{EqO}(x,y) \in \Gamma_O$*
- $V = \mathsf{EqRel}(V' \cup \{(\langle t, i\rangle, \langle t', i'\rangle)\}, \mathsf{Cells}(D))$, *for a candidate solution $\langle E, V' \rangle$ for (D, Σ) and $(t, t') \in q(D_{E,V})$ s.t. $q(x_t, y_t) \rightarrow \mathsf{EqV}(\langle x_t, i\rangle, \langle y_t, i'\rangle) \in \Gamma_V$.*

If additionally $D_{E,V} \models \Gamma_h^o \cup \Gamma_h^v \cup \Delta$, then we call $\langle E, V \rangle$ a solution *for (D, Σ). We use $\mathsf{Sol}(D, \Sigma)$ for the set of solutions for (D, Σ).*

Observe that by construction, every merge occurring in a solution can be justified by providing the sequence of rule applications and closure operations [11] that led to the merge being incorporated into the solution (alternatively, such steps may be presented as a proof tree, as formalized and illustrated in the long version of [56]). Importantly, as rule bodies do not involve any kind of negation (in particular, no $\neq$-atoms), 'later' merges cannot invalidate the reasons for performing an 'earlier' merge.

We note that a database-specification pair may admit zero, one, or several solutions. The absence of solutions arises from constraint violations (either initially present or introduced by the hard rules) which cannot be repaired solely through permitted merges. The existence of multiple solutions is due to some combinations of merges not being possible without violating the constraints, leading to a choice of which possible merges to include. We return to our running example[5] to illustrate solutions and the utility of local merges:

Example 2. *Starting from database D_{ex}, we can apply the soft rule σ_1^o to merge author ids a_1 and a_2 (more formally, we minimally extend the initial trivial equivalence relation E to include (a_1, a_2)). The resulting induced instance is obtained by replacing* all *occurrences of a_1 and a_2 by $\{a_1, a_2\}$. Note that the constraint δ_1 is now violated, since t_1 and t_2 match on aid, but have different names. If local merges were not permitted (as was the case in the original Lace framework), this would prevent (a_1, a_2) from belonging to any solution. However, thanks to the hard rule for values ρ_1^v, we can resolve this violation. Indeed, ρ_1^v is applicable and allows us to (locally) merge the names in facts t_1 and t_2. The new induced database contains {J. Smith, Joe Smith} in the name position of t_1 and t_2, but the names for t_3, t_4, t_5 remain as before. Note the importance of performing a local rather than a global merge: if we had grouped J. Smith with Joe Smith* everywhere*, this would force a merge of a_3 with a_4*

[5] Additional examples of Lace specifications and solutions can be found in [11, 12, 56, 57].

due to the hard rule ρ_1^o, which would in turn violate δ_2, again resulting in no solution containing (a_1, a_2). Following the local merge of the names of t_1 and t_2, the hard rule ρ_1^o becomes applicable and allows us (actually, forces us) to merge (globally) author ids a_1 and a_5. We let $\langle E_{\mathsf{ex}}, V_{\mathsf{ex}}\rangle$ be the equivalence relations obtained from the preceding rule applications. As the database induced by $\langle E_{\mathsf{ex}}, V_{\mathsf{ex}}\rangle$ satisfies all hard rules and constraints, $\langle E_{\mathsf{ex}}, V_{\mathsf{ex}}\rangle$ is a solution. Another solution is the pair of trivial equivalence relations, since the initial database D_{ex} satisfies the constraints and hard rules.

Rather than considering all solutions, it is natural to restrict attention to the 'best' ones. We shall therefore focus on solutions that are maximal w.r.t. set inclusion, i.e. derive as many merges as possible subject to the constraints. Alternative optimality criteria can also be considered, see Section 3.4.

DEFINITION 5. *The set* $\mathsf{MaxSol}(D, \Sigma)$ *of* maximal solutions *for* (D, Σ) *contains those* $\langle E, V\rangle \in \mathsf{Sol}(D, \Sigma)$ *for which there is no solution* $\langle E', V'\rangle \in \mathsf{Sol}(D, \Sigma)$ *with* $E \cup V \subsetneq E' \cup V'$.

EXAMPLE 3. *The solution* $\langle E_{\mathsf{ex}}, V_{\mathsf{ex}}\rangle$ *described in Example 2 is not maximal as the soft rule* σ_1^o *can be applied to get* (a_6, a_7) *or* (a_7, a_8). *Notice, however, that it is not possible to include both merges, otherwise by transitivity,* a_6, a_7, a_8 *would all be replaced by* $\{a_6, a_7, a_8\}$, *which would violate denial* δ_1 *due to paper* p_5. *We have two maximal solutions: the first extends* $\langle E_{\mathsf{ex}}, V_{\mathsf{ex}}\rangle$ *with* (a_6, a_7) *and the corresponding pair of names cells* $(\langle t_6, 2\rangle, \langle t_7, 2\rangle)$ *(due to* ρ_1^v*), and the second extends* $\langle E_{\mathsf{ex}}, V_{\mathsf{ex}}\rangle$ *with* (a_7, a_8) *and the corresponding name cells* $(\langle t_6, 2\rangle, \langle t_7, 2\rangle)$ *(via* ρ_1^v*).*

3.3 Properties of the Framework

We briefly highlight some interesting properties of the LACE framework.

Simulating Hard Rules We can show that hard rules can be simulated by soft rules in combination with denial constraints, provided that we allow denial constraints to use atoms with similarity relations. Specifically, a hard rule $\rho^o = \varphi(x, y, \mathbf{z}) \Rightarrow \mathsf{EqO}(x, y)$ can be replaced by a soft rule $\sigma_{\rho^o} = \varphi(x, y, \mathbf{z}) \dashrightarrow \mathsf{EqO}(x, y)$ and DC $\delta_{\rho^o} = \varphi(x, y, \mathbf{z}) \wedge x \neq y \rightarrow \bot$. Similarly, $\rho^v = \varphi(x_t, y_t, \mathbf{z}) \Rightarrow \mathsf{EqV}(\langle x_t, i\rangle, \langle y_t, j\rangle)$ is replaced by $\sigma_{\rho^v} = \varphi(x_t, y_t, \mathbf{z}) \dashrightarrow \mathsf{EqV}(\langle x_t, i\rangle, \langle y_t, j\rangle)$ and $\delta_{\rho^v} = \varphi(x_t, y_t, \mathbf{z}) \wedge u_{x_t,i} \neq u_{y_t,j} \rightarrow \bot$, with $u_{x_t,i}$ (resp. $u_{y_t,j}$) the ith (resp. jth) argument of the atom in $\varphi(x_t, y_t, \mathbf{z})$ with tid x_t (resp. y_t).

THEOREM 1. *Consider an ER specification* $\Sigma = \langle \Gamma_h^o \cup \Gamma_s^o, \Gamma_h^v \cup \Gamma_s^v, \Delta\rangle$ *over* $\mathcal{S}$, *and define* Σ' *as the specification* $\langle \Gamma^o_{h\rightsquigarrow s}, \Gamma^v_{h\rightsquigarrow s}, \Delta'\rangle$ *with* $\Gamma^o_{h\rightsquigarrow s} = \Gamma_s^o \cup \{\sigma_{\rho^o} \mid \rho^o \in \Gamma_h^o\}$, $\Gamma^v_{h\rightsquigarrow s} = \Gamma_s^v \cup \{\sigma_{\rho^v} \mid \rho^v \in \Gamma_h^v\}$, *and* $\Delta' = \Delta \cup \{\delta_\rho \mid \rho \in \Gamma_h^o \cup \Gamma_h^v\}$. *Then* $\mathsf{Sol}(D, \Sigma) = \mathsf{Sol}(D, \Sigma')$ *for each* $\mathcal{S}$*-database* D.

Local Can Simulate Global Interestingly, we can show that it is possible to simulate global merges of objects using local value merges. To formulate the result, we will use $\mathcal{S}_\mathsf{V}$ for the schema with the same relations as $\mathcal{S}$ but with all object positions changed to value positions, and use D_V for the dataset D but with all object constants treated as value constants.

THEOREM 2. *For every ER specification* $\Sigma = \langle \Gamma_O, \Gamma_V, \Delta\rangle$ *over schema* $\mathcal{S}$, *there exists a specification* $\Sigma' = \langle \emptyset, \Gamma'_V, \Delta\rangle$ *over* $\mathcal{S}_\mathsf{V}$ *s.t. for every* $\mathcal{S}$*-database* D: $\mathsf{Sol}(D_\mathsf{V}, \Sigma') = \{\langle \emptyset, V \cup V_E\rangle \mid \langle E, V\rangle \in \mathsf{Sol}(D, \Sigma)\}$, *where* $V_E = \{(\langle t, i\rangle, \langle t', j\rangle) \mid (t[i], t'[j]) \in E\}$.

PROOF SKETCH. Every rule $q(x, y) \rightarrow \mathsf{EqO}(x, y)$ is replaced by a rule $q(x_t, y_t) \rightarrow \mathsf{EqV}(\langle x_t, i\rangle, \langle y_t, j\rangle)$, where $\langle x_t, i\rangle$ (resp. $\langle y_t, j\rangle$) is any position that contains x (resp. y) in q. Additionally, we include all rules $P(x_t, \mathbf{u}_1^{i-1}, z, \mathbf{u}_{i+1}^k) \wedge P'(y_t, \mathbf{v}_1^{j-1}, z, \mathbf{v}_{j+1}^\ell) \Rightarrow \mathsf{EqV}(\langle x_t, i\rangle, \langle y_t, j\rangle)$ such that $\mathbf{u}_1^{i-1}$ abbreviates the tuple of distinct variables $u_1, \ldots, u_{i-1}$ (likewise for $\mathbf{u}_{i+1}^k$, $\mathbf{v}_1^{j-1}$, $\mathbf{v}_{j+1}^\ell$) and i and j are object positions of P/k and P'/ℓ w.r.t. the original schema $\mathcal{S}$. □

Note that there can be no analogous translation from local to global, for the simple reason that the equivalence relation for objects cannot distinguish between different occurrences of a same constant.

3.4 Alternative Optimality Criteria

While focusing on solutions with a $\subseteq$-maximal set of merges is arguably reasonable, there are other natural optimality criteria that can be used instead. For example, we may want to give more importance to a merge that is supported by multiple rules, or compare solutions based upon soft rule violations. To formalize these criteria, we will use the notion of active pair: (o, o') (resp. $(\langle t, i\rangle, \langle t', i'\rangle)$) is *active* in $D_{E,V}$ w.r.t. $q(x, y) \rightarrow \mathsf{EqO}(x, y)$ (resp. $q(x_t, y_t) \rightarrow \mathsf{EqV}(\langle x_t, i\rangle, \langle y_t, i'\rangle)$) if $(o, o') \in q(D_{E,V})$ (resp. $(t, t') \in D_{E,V}$). We then define $\mathbf{ap}(D, E, V, \Gamma)$ as the set of all (μ, ρ) such that pair μ is active in $D_{E,V}$ w.r.t. rule $\rho \in \Gamma$. Our proposed optimality criteria are obtained by associating each solution $\langle E, V\rangle$ with one of the following sets (using Γ for $\Gamma_O \cup \Gamma_V$):

$$\begin{aligned}
\text{EQ}(E, V) &= E \cup V\\
\text{SP}(E, V) &= \{(\mu, \rho) \in \mathbf{ap}(D, E, V, \Gamma) \mid \mu \in E \cup V\}\\
\text{AB}(E, V) &= \{\mu \mid (\mu, \rho) \in \mathbf{ap}(D, E, V, \Gamma), \mu \notin E \cup V\}\\
\text{VIO}(E, V) &= \{(\mu, \rho) \in \mathbf{ap}(D, E, V, \Gamma) \mid \mu \notin E \cup V\}
\end{aligned}$$

Observe that $\text{SP}(E,V)$ refines $\text{EQ}(E,V)$ by indicating the supporting rules for merges. Likewise, $\text{AB}(E,V)$ gives only the active but absent pairs, while $\text{VIO}(E,V)$ records which soft rules the absent pair violates. The resulting optimality criteria are:

- maxES/maxEC: maximize $\text{EQ}(E,V)$
- maxSS/maxSC: maximize $\text{SP}(E,V)$
- minAS/minAC: minimize $\text{AB}(E,V)$
- minVS/minVC: minimize $\text{VIO}(E,V)$

where the final S (resp. C) indicates comparison using set inclusion (resp. set cardinality). For example, a solution $\langle E,V\rangle$ is minVC-optimal if there is no other solution $\langle E',V'\rangle$ such that $\text{VIO}(E',V') < \text{VIO}(E,V)$. Note that maxES-optimal solutions correspond to the maximal solutions of Definition 5. It can be shown that the optimality criteria give rise to different sets of optimal solutions, except for maxES and maxSS, which actually coincide. Thus, there are overall seven distinct optimality criteria.

4 Complexity Results

In this section, we analyze the computational complexity of the main decision problems associated with the framework. Since the database size is typically order of magnitude larger than the other components, we focus on the *data complexity* measure, i.e. the complexity w.r.t. the size of the database D (and also $\langle E,V\rangle$ for those problems that require it).

We start with the solution recognition problem (Rec): decide if a given $\langle E,V\rangle$ belongs to $\mathsf{Sol}(D,\Sigma)$. To solve this problem, it is enough to verify that $D_{E,V} \models \Gamma_h^o \cup \Gamma_h^v \cup \Delta$ and that $\langle E,V\rangle$ is a candidate solution for (D,Σ). The latter can be done by starting with the empty set of merges and then repeatedly applying the rules in Σ to check whether some $\alpha \in E \cup V$ can be added to the current set.

Theorem 3. *Rec is P-complete.*

By contrast, the problem of deciding whether there exists a solution (Existence) is intractable.

Theorem 4. *Existence is NP-complete.*

Proof sketch. The upper bound is trivial: guess $\langle E,V\rangle$ and check if it is a solution for (D,Σ).

The lower bound is by reduction from the 3CNF problem. Consider $\phi = c_1 \wedge \ldots \wedge c_m$ over the variables $x_1,\ldots,x_n$, where $c_i = \ell_{i,1} \vee \ell_{i,2} \vee \ell_{i,3}$. Denote by $x_{i,j}$ the variable of $\ell_{i,j}$, and set $s_{i,j} = t$ if $\ell_{i,j} = x_{i,j}$ and $s_{i,j} = f$ if $\ell_{i,j} = \neg x_{i,j}$. We encode ϕ with a database comprising the following facts:

$$\{V(t_{x_i}, x_i) \mid 1 \le i \le n\} \cup$$
$$\{Prec(t_{p_i}, x_i, x_{i+1}) \mid 1 \le i < n\} \cup$$
$$\{R_{s_{i,1}s_{i,2}s_{i,3}}(t_{c_i}, x_{i,1}, x_{i,2}, x_{i,3}) \mid 1 \le i \le m\} \cup$$
$$\{FV(t_f, x_1), LV(t_l, x_n), T(t_1, 1), F(t_0, 0)\} \cup$$
$$\{Q(t_{Q_0}, 0), Q(t_{Q_1}, 1), C_1(t_{C_1}, c_1), C_2(t_{C_2}, c_2)\}.$$

For instance, a clause of the form $c_i = x_k \vee \neg x_z \vee x_w$ is represented as $R_{tft}(t_{c_i}, x_k, x_z, x_w)$. The fixed ER specification Σ_{3CNF} has soft rules $V(t_1,x) \wedge Q(t_2,y) \wedge FV(t_3,x) \dashrightarrow \mathsf{EqO}(x,y)$, $V(t_1,x) \wedge Q(t_2,y) \wedge Prec(t_3,x_p,x) \wedge Q(t_4,x_p) \dashrightarrow \mathsf{EqO}(x,y)$, and $C_1(t_1,x) \wedge C_2(t_2,y) \wedge Q(t_3,z) \wedge LV(t_4,z) \dashrightarrow \mathsf{EqO}(x,y)$. The first two allow variables to merge with either 0 or 1. Once every variable has been assigned a truth value, the third rule merges c_1 and c_2. The DCs in Σ_{3CNF} ensure the merges yield a proper truth assignment ($F(t_1,y) \wedge T(t_2,y) \rightarrow \bot$) not violating any clause ($R_{tft}(t_1,y_1,y_2,y_3) \wedge F(t_2,y_1) \wedge T(t_3,y_2) \wedge F(t_4,y_3) \rightarrow \bot$, and similarly for other clause types). Finally, $C_1(t_1,y_1) \wedge C_2(t_2,y_2) \wedge y_1 \neq y_2 \rightarrow \bot$ requires c_1 and c_2 to be merged, which means a truth assignment is generated. □

Another central problem is deciding whether $\langle E,V\rangle \in \mathsf{MaxSol}(D,\Sigma)$ (MaxRec). The coNP upper bound is easy (guess $\langle E',V'\rangle$, check that $\langle E',V'\rangle \in \mathsf{Sol}(D,\Sigma)$ and $E \cup V \subsetneq E' \cup V'$), and a coNP lower bound can be obtained by slightly extending the one for Existence. The basic idea is to introduce a new fact $C(t^*,e,e')$ and new soft rule $C(t,x,y) \rightarrow \mathsf{EqO}(x,y)$. The first soft rule is then replaced by $V(t_1,x) \wedge Q(t_2,y) \wedge FV(t_3,x) \wedge C(t_4,z,z) \dashrightarrow \mathsf{EqO}(x,y)$, allowing x_1 to merge with either 0 or 1 (and enabling such merges for later variables) only if e and e' have been previously merged. As a result, $\langle \mathsf{EqRel}(\emptyset, \mathsf{Obj}(D^\phi)), \mathsf{EqRel}(\emptyset, \mathsf{Cells}(D^\phi))\rangle$ is a maximal solution for $(D^\phi, \Sigma'_{\text{3CNF}})$ iff ϕ is unsatisfiable.

Theorem 5. *MaxRec is coNP-complete.*

Other key tasks involve formal reasoning over the maximal solutions of a database-specification pair. We start with two tasks that enable a *credulous* form of reasoning: deciding whether (i) a given merge α is such that $\alpha \in \langle E,V\rangle$ for *some* $\langle E,V\rangle \in \mathsf{MaxSol}(D,\Sigma)$ (PossMerge) and (ii) a given CQ q and tuple $\mathbf{c}$ is such that $\mathbf{c} \in q(D_{E,V})$ for *some* $\langle E,V\rangle \in \mathsf{MaxSol}(D,\Sigma)$ (PossAns).

Theorem 6. *Both PossMerge and PossAns are NP-complete.*

Proof sketch. The upper bounds are based on guessing $\langle E,V\rangle$ and then checking that $\langle E,V\rangle \in$

$\mathsf{Sol}(D,\Sigma)$ and $\alpha \in E \cup V$ (resp. $\mathbf{c} \in q(D_{E,V})$). Obviously, if $\alpha \in \langle E,V\rangle$ (resp. $\mathbf{c} \in q(D_{E,V})$) for $\langle E,V\rangle \in \mathsf{Sol}(D,\Sigma)$, then $\alpha \in \langle \mathcal{E},\mathcal{V}\rangle$ (resp. $\mathbf{c} \in q(D_{\mathcal{E},\mathcal{V}})$) for some $\langle \mathcal{E},\mathcal{V}\rangle \in \mathsf{MaxSol}(D,\Sigma)$ with $E \cup V \subsetneq \mathcal{E} \cup \mathcal{V}$. For the lower bounds, consider the specification obtained from the one in the proof of Theorem 4 by removing the last denial constraint. Then, (c_1, c_2) is a possible merge (resp. (c_1) is a possible answer to $q(x) = C_1(x) \wedge C_2(x)$) iff ϕ is satisfiable. □

We now investigate a *skeptical* form of reasoning through the decision problems CertMerge and CertAns, defined as PossMerge and PossAns, respectively, but with the additional requirement that $\alpha \in \langle E,V\rangle$ for *all* $\langle E,V\rangle \in \mathsf{MaxSol}(D,\Sigma)$ in the case of CertMerge, and that $\mathbf{c} \in q(D_{E,V})$ for *all* $\langle E,V\rangle \in \mathsf{MaxSol}(D,\Sigma)$ in the case of CertAns. This problem can be solved by guessing $\langle E,V\rangle$, and then checking $\langle E,V\rangle \in \mathsf{MaxSol}(D,\Sigma)$ and $\alpha \notin E \cup V$ (resp. $\mathbf{c} \notin q(D_{E,V})$), which puts the problems in Π_2^p. The main difference with PossMerge and PossAns is that one needs to check that the guessed pair is a *maximal* solution, not just a solution.

Theorem 7. *Both CertMerge and CertAns are Π_2^p-complete.*

All of our lower bounds hold already for fixed ER specifications having only rules for objects, and, moreover, they can be adapted to apply to ER specifications using only FDs as denial constraints.

We now consider the case of *restricted* specifications, i.e. ER specifications whose DCs do not use inequality atoms. Such $\neq$-free DCs are widely used in ontologies, e.g. to express class disjointness, and are available in popular ontology languages such as *DL-Lite* [19]. While the results in Theorems 3 and 6 apply already to restricted ER specifications, the other tasks become easier under standard complexity assumptions. Intuitively, this is because constraint violations are preserved under merges.

Theorem 8. *For restricted ER specifications, Existence and MaxRec are P-complete, while CertMerge and CertAns are coNP-complete.*

For the other optimality criteria discussed in Section 3.4, we studied the complexity of recognizing optimal solutions [57]. The next result shows that this problem is coNP-complete for all the optimality criteria, just as in the case of maxES (Theorem 5).

Theorem 9. *For all defined optimality criteria, recognition of optimal solutions is coNP-complete.*

Interestingly, for restricted ER specifications, while the problem is P-complete for the optimality criteria based on set-inclusion (as for maxES, see Theorem 8), the problem remains intractable for the optimality criteria based on set cardinality.

Theorem 10. *For restricted ER specifications, recognition of optimal solutions is P-complete (resp. coNP-complete) for all the optimality criteria based on set inclusion (resp. set cardinality).*

5 LACE Implementation

In order to explore the practical interest of Lace, we have implemented the framework using answer set programming (ASP), a well-studied paradigm for declarative problem solving [43]. The suitability of employing ASP for tackling data quality tasks has been previously demonstrated by work on data cleaning with (relational) MDs [4, 5] and consistent query answering [25, 45].

5.1 ASP Encoding of LACE specifications

Given an ER specification Σ, the ASP encoding of Σ is a program Π_Σ containing an ASP rule for each (hard or soft) rule in Σ. Predicates `eqo` and `eqv` are used to store merges of objects and values respectively, and additional rules are used to ensure that `eqo` and `eqv` are equivalence relations.

Rather than providing the full encoding, which requires introducing quite a lot of notation, we shall describe the main ideas underlying the encoding. A hard rule for objects will have head atom $\texttt{eqo}(X,Y)$, enforcing that X and Y are merged. A soft rule for objects can be elegantly encoded using a choice rule (in ASP parlance) whose head $\{\texttt{eqo}(X,Y)\}$ allows but does not require that $\texttt{eqo}(X,Y)$ is made true when the rule body holds. Rules for values are instead use head atom $\texttt{eqv}(T,I,T',J)$ to encode merges of cells, again with choice rules used to capture the semantics of soft rules for values.

The translation of Lace rule bodies into the corresponding ASP rule bodies is a bit more involved, as we need to simulate the effect of evaluating the rule body over the induced instance. This is accomplished by instantiating every variable position with a distinct variable, then using adding `eqo` and `eqv` to enforce that the required positions 'join' in the induced database. The encoding of rule bodies must also ensure that similarity atoms are evaluated using the common set of values for the compared variables (cf. Definition 3).

As the following result shows, the stable models of the ASP program capture Lace solutions:

Theorem 11. *For every database D and specification $\Sigma = \langle \Gamma_O, \Gamma_V, \Delta\rangle$: $\langle E,V\rangle \in \mathsf{Sol}(D,\Sigma)$ iff $E = \{(a,b) \mid \texttt{eqo}(a,b) \in M\}$ and $V =$*

$\{(\langle t, i\rangle, \langle t', i'\rangle) \mid \texttt{eqv}(t, i, t', i') \in M\}$ *for a stable model* M *of* (Π_Σ, D).

5.2 Similarity Computation

Differently from the original ASP encoding in [11], similarity relations are implemented with a predicate $\texttt{sim}_i(X, Y, S)$, where X and Y are instantiated with the constants to be assessed for similarity, and S with a similarity score. This allows the same similarity measure to be used in different rules with different thresholds (specified using comparison atoms, e.g. $S > 95$, in the rule body). In the specifications used in our evaluation, we employ similarity atoms based upon Levenshtein distance for numerical constants, Jaro–Winkler distance for short strings, and TF–IDF cosine score for long textual values.

A key challenge is how to efficiently evaluate the $\texttt{sim}_i$ atoms, since the input dataset does not contain $\texttt{sim}_i$ facts, which must instead be computed via external functions (e.g., string similarity measures). A naïve approach is to precompute the set of all facts $\texttt{sim}_i(c, d, s)$ where c and d are compatible values and $s = f_i(c, d)$, with f_i the function underlying the relation $\texttt{sim}_i$. Although this only needs polynomially many function calls in $|D|$, it is prohibitively costly on even moderately-sized databases.

This led us to explore a more sophisticated strategy for similarity computation [56], which exploits the program structure to better identify which pairs of facts need to be compared. Roughly speaking, this is achieved by considering different simplifications of the original program, coupled with online calls to the external functions. The empirical evaluation conducted in [56] shows that this strategy can achieve substantial improvements in runtime and memory efficiency, especially on larger datasets.

5.3 ASPEN and ASPEN$^+$

The systems ASPen [56] and ASPen$^+$ [57] implement the Lace framework, employing the ASP encoding described in Section 5.1 and exploiting the capabilities of the `clingo` ASP solver [36] to generate and reason about ER solutions. By utilizing the diverse reasoning modes available in `clingo`, ASPen can produce one or more (maximal) solutions as well as other relevant merge sets, like the set of possible merges and an approximation of the set of certain merges. Explainability is achieved through integration with `xclingo` [18], which enables the generation of proof trees that justify individual merges by explicitly tracing the rule applications that led to a merge being derived.

ASPen$^+$ extends ASPen's functionality by introducing support for local merges, thereby enabling context-specific value resolution in addition to global merges. Furthermore, ASPen$^+$ implements the set of optimality criteria described in Section 3.4, allowing the selection of preferred solutions according to various optimality criteria. This optimization capability is realized through the `asprin` framework [15, 16], which provides declarative preference handling within ASP.

5.4 Experiments

ASPen and ASPen$^+$ are publicly available[6], and experimental results show strong performance across real-world ER benchmarks and synthetic data, each with ground truth merges. Both systems have been evaluated against existing open-source systems, Magellan and JedAI [38, 48], that support rule-based ER, which we take as our baselines.

Effectiveness and Scalability Across all datasets, ASPen and ASPen$^+$ achieved consistently higher F1 scores than the baseline systems, with performance gaps of up to 86% on multi-relational datasets where traditional ER approaches perform poorly. These quality gains, however, came with a cost in computational efficiency: both baselines were significantly faster than the ASP-based ones.

The scalability analysis indicates that runtime is sensitive to several factors. Increasing the *data size* or the *duplicate ratio*, or reducing the *similarity thresholds*, leads to substantial slowdowns. In extreme cases, these variations caused up to a 300-fold increase in execution time. This reflects the higher computational complexity of the ASP-based approach, particularly under parameter settings that expand the search space of possible merges.

Impact of Local Merges With global semantics alone, ASPen programs often admit no solution when all natural FDs were included in specifications, especially in noisy datasets. ASPen$^+$, by supporting local semantics, accommodated all FDs and achieved higher-quality results, even outperforming ASPen in cases where both could satisfy the constraints. Overall, this analysis shows that local semantics provides greater robustness to data variability and constraint interactions, enabling more complete and accurate ER solutions.

Optimality Criteria On datasets with few null values and little variation in values, criteria that maximize the number of merges (maxE and maxS) achieved the highest F1 scores because they focus on coverage. On noisier datasets, criteria that minimize rule violations (minA and minV) performed

[6] https://github.com/zl-xiang/Aspen

better, as their emphasis on precision reduced incorrect merges. In all settings, solutions based on set inclusion were usually faster to compute, while solutions based on the number of merges were often closer to the gold standard. The improvement in quality from using the latter often required significantly higher computation times, making the former more suitable when resources are limited.

6 Combining ER with Repairs

Real-world databases may suffer from multiple data quality issues. Some constraint violations may result from the use of different constants for the same entity, and thus may be resolved through merging constants, but others may stem from the presence of erroneous facts and can only be resolved by repairing the data, i.e. removing or modifying facts. A pipeline approach, applying ER and repairing methods in sequence, may miss useful synergies. For example, by merging two constants, we may resolve an FD violation without the need to delete facts, while conversely, deleting incorrect facts may enable some desirable merges. This suggests the interest of developing holistic approaches to jointly deduplicating and repairing data, an idea which was been advocated in [23, 34] but remains little explored.

These considerations motivated us to propose the Replace framework [12], an extension of Lace that allows for both merge and fact deletion operations. Fact deletions make it possible to obtain meaningful solutions when $\mathsf{Sol}(D, \Sigma) = \emptyset$, but also to discover additional merges that were blocked due to constraint violations. The Replace framework employs the same form of specifications as Lace, but redefines the notion of solution, which now takes the form[7] $\langle R, E, V\rangle$, with E, V equivalence relations as before and R is a set of facts to delete from D.

Definition 6. *Given a specification Σ over $\mathcal{S}$ and $\mathcal{S}$-database D, we call $\langle R, E, V\rangle$ a* Rep-solution for (D, Σ) *if $R \subseteq D$ and $\langle E, V\rangle \in \mathsf{Sol}(D \setminus R, \Sigma)$.*

Similarly to Lace, we naturally prefer solutions that contain more merges. However, we also want to retain as much information as possible, hence should minimize fact deletions, as is done when defining repairs. These two criteria may conflict, as deleting more facts may enable more merges. This lead us to consider three natural ways to compare Rep-solutions: give priority to maximizing merges (Mer), give priority to minimizing deletions (Del), or adopt the Pareto principle and accord equal priority to both criteria (Par). Using $X \in \{\text{Mer}, \text{Del}, \text{Par}\}$ for comparison, we obtain the set $\mathsf{Sol}_X^{\text{Rep}}(D, \Sigma)$ of $\preceq_X$-optimal Rep-solutions.

It is easily verified that we always have $\mathsf{Sol}_{\text{Mer}}^{\text{Rep}}(D, \Sigma) \subseteq \mathsf{Sol}_{\text{Par}}^{\text{Rep}}(D, \Sigma)$ and $\mathsf{Sol}_{\text{Del}}^{\text{Rep}}(D, \Sigma) \subseteq \mathsf{Sol}_{\text{Par}}^{\text{Rep}}(D, \Sigma)$, while the converse inclusions do not hold in general. We further observe that $\langle \emptyset, E, V\rangle \in \mathsf{Sol}_{\text{Del}}^{\text{Rep}}(D, \Sigma)$ iff $\langle \emptyset, E, V\rangle \in \mathsf{Sol}_{\text{Par}}^{\text{Rep}}(D, \Sigma)$ iff $\langle E, V\rangle \in \mathsf{MaxSol}(D, \Sigma)$. Thus, maximal solutions in Lace are special cases of $\preceq_{\text{Del}}$- and $\preceq_{\text{Par}}$-optimal solutions (an analogous property does not hold for $\preceq_{\text{Mer}}$). Rep-solutions can also be related to the subset repairs employed in consistent query answering [3, 7, 20]. Indeed, if we consider (D, Σ) with $\Sigma = \langle \emptyset, \emptyset, \Delta\rangle$, then $\mathsf{Sol}_{\text{Mer}}^{\text{Rep}}(D, \Sigma)$, $\mathsf{Sol}_{\text{Del}}^{\text{Rep}}(D, \Sigma)$, and $\mathsf{Sol}_{\text{Par}}^{\text{Rep}}(D, \Sigma)$ coincide and contain only solutions of the form $(R, \mathsf{trivE}, \mathsf{trivCells})$, with trivE and $\mathsf{trivCells}$ the trivial equivalence relations over $\mathsf{Obj}(D)$ and $\mathsf{Cells}(D)$. It is readily verified that $\langle R, \mathsf{trivE}, \mathsf{trivCells}\rangle \in \mathsf{Sol}_{\text{Mer}}^{\text{Rep}}(D, \Sigma) = \mathsf{Sol}_{\text{Del}}^{\text{Rep}}(D, \Sigma) = \mathsf{Sol}_{\text{Par}}^{\text{Rep}}(D, \Sigma)$ iff $D \setminus R$ is a repair.

The complexity analysis of Replace carried out in [12] reveals that in almost all cases, the addition of delete operations to Lace does not affect the complexity of recognizing (maximal / optimal) solutions or certain and possible answers.

7 Overview of Logic-Based ER Methods

As a foundational and multifaceted task in computer science, entity resolution has been tackled using a variety of different approaches, including probabilistic models, (deep) learning techniques, and logical methods [21]. In this section, we provide a brief overview of logic-based approaches to ER and related problems[8]. We will compare these works by considering (i) which ER problem is tackled and what constitutes a solution, (ii) what kind of rules and/or constraints are employed, (iii) the nature of the semantics (static vs. dynamic, local and/or global merges), and (iv) the existence of an accompanying implementation.

Dedupalog [2] was the first logic-based framework targeting collective ER. It employs soft and hard datalog-style rules and also allows for rules with negated heads, to indicate likely non-merges. Dedupalog allows for recursive rules, but due to the static semantics, it is unclear how to extract a non-circular derivation of produced merges. The semantics can be characterized as global, as solutions in Dedu-

[7] As Replace [12] extends the original Lace framework [11], it only supports global merges. Definition 6 adapts the notion of solution from [12] to accommodate local merges. It can be verified that the complexity results in [12] hold also for this modified definition.

[8] A detailed account of early logic-based approaches and their precursors can be found in Chapter 4 of [29].

palog define equivalence relations over entity references from designated relations. Solutions are required to satisfy all hard rules and should minimize violations of soft rules. However, for efficiency reasons, the Dedupalog system (not publicly available) generates a single approximately optimal solution.

Matching dependencies (MDs) specify conditions under which pairs of attribute values in database facts must be matched [27, 29, 32], i.e., made equal. Formally, an MD is an expression of the form

$$R_1[\vec{X}_1] \approx R_2[\vec{X}_2] \;\; \rightarrow \;\; R_1[Y_1] \doteq R_2[Y_2],$$

which states that if the projections of an R_1-fact t_1 and an R_2-fact t_2 onto attributes $\vec{X}_1$ and $\vec{X}_2$ are pairwise similar, then the Y_1-value of t_1 and the Y_2-value of t_2 must be made equal. Relational MDs [4, 5] generalize MDs by allowing additional atoms in the body, supporting collective scenarios. (Relational) MDs are equipped with a dynamic semantics: when the body condition is satisfied, values are (locally) updated to ensure satisfaction of the head. In [8], this is formalized using a chase-like procedure that repairs violations of MDs, using matching functions to determine the resulting value when two values are matched (rather than using the set of merged constants). Although (relational) MDs can be viewed as hard constraints (since they must be satisfied), the order in which rules are applied affects the outcome, as value modifications may lead to multiple possible solutions.

More recently, entity-enhancing rules (REEs) have been introduced [33], which extend both relational MDs and (conditional) functional dependencies by incorporating machine learning predicates and attribute-value comparisons. In the context of entity resolution, REEs focus on the global matching of tuple IDs through a chase-like procedure, which if successful, yields a unique updated data instance in which all REEs are satisfied (REEs are thus treated as hard rules). Although the framework can infer (in)equalities among tuple cells, the presence of multiple representations of a data value is regarded as an error that must be resolved by the end user. Aside from entity resolution, REEs can also be used for other data quality tasks, like conflict resolution and data imputation.

The recently proposed CERQ framework [26] considers ER in the setting of knowledge bases consisting of facts, tuple-generating dependencies (tgds), and equality-generating dependencies (egds). Intuitively, the egds act as hard rules for objects and values, and the tgds support (open-world) inference of new facts. The chase-based semantics is dynamic and supports both local and global merges, where the notion of instance takes a very similar form to our notion of induced database (having, in particular, sets of values in value positions). Interestingly, although developed independently, the semantics for the satisfaction of queries and rules over instances with sets of constants shares the same principle adopted in Lace. As the CERQ framework does not consider soft rules and denial constraints, it is possible to define the notion of universal solution as the preferred output (which can be used to support conjunctive query answering when the chase terminates). Finally, we mention that the CERQ framework does not have an implementation.

A declarative framework for entity linking (EL) based upon link-to-source constraints was presented in [17]. In contrast to ER, whose aim is to infer which entities that correspond to the same real-world object, this work is concerned with discovering other kinds of binary relations linking pairs of entities. As a result, link relations are not constrained to equivalence relations, and it is not evident how one can force relations to act as equivalence relations to adapt the framework to handle collective ER (in particular, recursive ER scenarios, cf. discussion in [11]). The static semantics characterizes a space of maximal solutions, from which notions of certain and possible links are defined.

There have also been several efforts to develop practical systems for ER based upon declarative formalisms. Prominent examples include the open-source systems Magellan [38] and JeDAI [48], which address simpler ER settings that match tuples from a pair of tables (or within a single table). Both systems support syntactically simple rules, formulated as single-pass conditions based on similarity measures between attribute values of tuple pairs. ERBlox [5] is a system for collective ER based upon relational MDs. It represents one of the earliest efforts to combine machine learning techniques with declarative approaches to ER. In particular, ERBlox employs ML techniques to construct a classifier that identifies blocks of duplicate candidates, over which MDs are subsequently applied for entity resolution. More recently, building on the REE framework, the industrial-scale data-cleaning system Rock [6] has been developed to address a range of data quality issues, including collective ER. These systems address scalability challenges through blocking strategies and/or parallelization and offer further functionalities to simplify usage, e.g. user interfaces, default settings, debugging, use of external KBs, support for semi-structured or unstructured data. They also showcase the interest of combining ML and declarative approaches.

8 Perspectives

We have briefly surveyed recent developments in logic-based entity resolution, as exemplified by the Lace framework. Key foundational and conceptual advances include: a differentiated treatment of object and value merges using global and local semantics, the use of dynamic semantics to enable justifiability in the presence of recursive rules, and the consideration of a space of (optimal) solutions, which can be explored using certain and possible merges and query answers. Valuable insights have also been gained from experimenting with implementations of logic-based ER, underscoring the importance of dedicated optimisations and the interest of combining logical and ML methods. Despite these important advances, many interesting foundational and practical questions remain to be tackled. We mention a few items high on our agenda:

Representation of Query Answers One of the original motivations for developing Lace and its successor Replace was to be able to evaluate queries w.r.t. a space of (Rep-)solutions, in the spirit of consistent query answering. Note however that it is not at all obvious how best to present query results in a way that makes clear which constants have been merged and avoids returning distinct yet equivalent answer tuples. For example, if we pose the query $\exists v.P(v, x, y)$ to a database containing $P(t, c_1, c_2)$ and $P(t', c_3, c_4)$, with (c_1, c_3) and (c_2, c_4) certain object merges, then we get four certain answers: (c_1, c_2), (c_1, c_4), (c_3, c_2), and (c_3, c_4). However, we would tempted to return instead a single answer tuple consisting of *sets of constants*: $(\{c_1, c_3\}, \{c_2, c_4\})$. This idea motivated the notion of *most informative (certain or possible) answers* [12], but the definition only handles global merges and lacks a practical algorithm.

Different Forms of Explanations The notions of justification and proof trees that have been defined for Lace [11, 56] can be used to explain to users how a given merge was obtained in a given solution. It would be interesting however to consider additional forms of explanations that concern the whole space of (maximal or optimal) solutions. For example, how can we justify why a given merge (or answer) is certain, or why a possible merge (or answer) is *not* certain? Some first ideas for how to formalize such explanations might be gleaned from prior work on explaining query (non)answers under repair-based semantics [10].

Integration with Ontologies It would also be relevant to extend Lace to the setting of ontology-based data access [51, 58], in which an ontology is used to provide a convenient vocabulary for query formulation and to specify domain knowledge, which can be exploited when answering queries. To the best of our knowledge, the only work that has considered entity resolution in the context of ontologies is the recent work on CERQ [26]. However, it is non-trivial to incorporate soft ER rules and denial constraints into the latter framework. Moreover, there are other interesting questions to explore, such as how to accommodate mappings that link the data to the ontology and which may involve the creation of new entity-referring constants.

Implementation of Holistic Approaches We plan to build upon ASPen to develop an implementation of the Replace framework, also drawing inspiration from existing ASP-based implementations of consistent query answering [25, 45]. The computation of certain answers to queries, which has not yet been incorporated into ASPen, is an important functionality that will require care to implement due to its higher complexity (Π_2^p).

Scalability While our experiments show that ASPen can successfully handle some real-world ER scenarios, scalability remains an issue. We plan to explore the potential of employing specialized data structures or custom procedures for handling equivalence relations, as has been considered for Datalog reasoners [46, 52]. As parallelization has been successfully employed in some rule-based ER systems [24], another promising direction is developing parallel algorithms, building on prior work in parallel Datalog reasoning [1, 50] and ASP solving [37].

Learning ER Specifications A major barrier to adopting logic-based approaches is the difficulty of obtaining accurate ER rules. Most existing work on rule learning for ER targets single-pass matching rules within one or two tables [39, 40, 53], although there has been some relevant recent research on discovering entity-enhancing rules [30,31]. Moreover, the related questions of learning conjunctive queries [14, 55] and mining database constraints [22, 44, 49] have also been extensively investigated. In addition to learning ER specifications, it is also interesting to continue to explore the use of machine-learning methods for tuning specifications (e.g. setting the score thresholds for similarity relations) and for obtaining more informative similarity measures (cf. use of ML predicates in [24]).

Acknowledgments

This work has been supported by the ANR AI Chair INTENDED (ANR-19-CHIA-0014) and by MUR under the PNRR project FAIR (PE0000013).

References

[1] T. Ajileye and B. Motik. Materialisation and data partitioning algorithms for distributed RDF systems. *J. Web Semant.*, 2022.

[2] A. Arasu, C. Ré, and D. Suciu. Large-scale deduplication with constraints using dedupalog. In *Proc. of ICDE*, 2009.

[3] M. Arenas, L. E. Bertossi, and J. Chomicki. Consistent query answers in inconsistent databases. In *Proc. of PODS*, 1999.

[4] Z. Bahmani and L. E. Bertossi. Enforcing relational matching dependencies with datalog for entity resolution. In *Proc. of FLAIRS*, 2017.

[5] Z. Bahmani, L. E. Bertossi, and N. Vasiloglou. Erblox: Combining matching dependencies with machine learning for entity resolution. *Int. J. Approx. Reason.*, 2017.

[6] Z. Bao, B. Binbin, W. Fan, D. Li, M. Li, K. Lin, W. Lin, P. Liu, P. Liu, Z. Lv, M. Ouyang, C. Sun, S. Tang, Y. Wang, Q. Wei, X. Wu, M. Xie, J. Zhang, R. Zhao, J. Zhu, and Y. Zhu. Rock: Cleaning data with both ML and logic rules. *Proc. VLDB Endow.*, 2024.

[7] L. E. Bertossi. *Database Repairing and Consistent Query Answering*. Morgan & Claypool Publishers, 2011.

[8] L. E. Bertossi, S. Kolahi, and L. V. S. Lakshmanan. Data cleaning and query answering with matching dependencies and matching functions. *Theory Comput. Syst.*, 2013.

[9] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. *ACM Trans. Knowl. Discov. Data*, 2007.

[10] M. Bienvenu, C. Bourgaux, and F. Goasdoué. Computing and explaining query answers over inconsistent dl-lite knowledge bases. *J. Artif. Intell. Res.*, 2019.

[11] M. Bienvenu, G. Cima, and V. Gutiérrez-Basulto. LACE: A logical approach to collective entity resolution. In *Proc. of PODS*, 2022.

[12] M. Bienvenu, G. Cima, and V. Gutiérrez-Basulto. REPLACE: A logical framework for combining collective entity resolution and repairing. In *Proc. of IJCAI*, 2023. Long version available at `https://orca.cardiff.ac.uk/id/eprint/159626/1/main.pdf`.

[13] M. Bienvenu, G. Cima, V. Gutiérrez-Basulto, and Y. Ibáñez-García. Combining global and local merges in logic-based entity resolution. In *Proc. of KR*, 2023. Long version available at `https://arxiv.org/pdf/2305.16926`.

[14] A. Bonifati, R. Ciucanu, and S. Staworko. Learning join queries from user examples. *ACM Trans. Database Syst.*, 2016.

[15] G. Brewka, J. P. Delgrande, J. Romero, and T. Schaub. asprin: Customizing answer set preferences without a headache. In *Proc. of AAAI*, 2015.

[16] G. Brewka, J. P. Delgrande, J. Romero, and T. Schaub. A general framework for preferences in answer set programming. *Artif. Intell.*, 2023.

[17] D. Burdick, R. Fagin, P. G. Kolaitis, L. Popa, and W. Tan. A declarative framework for linking entities. *ACM Trans. Database Syst.*, 2016.

[18] P. Cabalar, J. Fandinno, and B. Muñiz. A system for explainable answer set programming. In *Proc. of ICLP*, 2020.

[19] D. Calvanese, G. D. Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. *J. Autom. Reason.*, 2007.

[20] J. Chomicki and J. Marcinkowski. Minimal-change integrity maintenance using tuple deletions. *Inf. Comput.*, 2005.

[21] V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, and K. Stefanidis. An overview of end-to-end entity resolution for big data. *ACM Comput. Surv.*, 2021.

[22] X. Chu, I. F. Ilyas, and P. Papotti. Discovering denial constraints. *Proc. VLDB Endow.*, 2013.

[23] X. Chu, I. F. Ilyas, and P. Papotti. Holistic data cleaning: Putting violations into context. In *Proc. of ICDE*, 2013.

[24] T. Deng, W. Fan, P. Lu, X. Luo, X. Zhu, and W. An. Deep and collective entity resolution in parallel. In *Proc. of ICDE*, 2022.

[25] T. Eiter, M. Fink, G. Greco, and D. Lembo. Repair localization for query answering from inconsistent databases. *ACM Trans. Database Syst.*, 2008.

[26] R. Fagin, P. G. Kolaitis, D. Lembo, L. Popa, and F. Scafoglieri. A framework for combining entity resolution and query answering in knowledge bases. In *Proc. of KR*, 2023.

[27] W. Fan. Dependencies revisited for improving data quality. In *Proc. of PODS*, 2008.

[28] W. Fan and F. Geerts. *Foundations of Data Quality Management*. Morgan & Claypool

Publishers, 2012.
[29] W. Fan and F. Geerts. *Foundations of Data Quality Management*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2012.
[30] W. Fan, Z. Han, Y. Wang, and M. Xie. Parallel rule discovery from large datasets by sampling. In *Proc. of SIGMOD*, 2022.
[31] W. Fan, Z. Han, M. Xie, and G. Zhang. Discovering top-k relevant and diversified rules. *Proc. ACM Manag. Data*, 2024.
[32] W. Fan, X. Jia, J. Li, and S. Ma. Reasoning about record matching rules. *Proc. VLDB Endow.*, 2009.
[33] W. Fan, P. Lu, and C. Tian. Unifying logic rules and machine learning for entity enhancing. *Sci. China Inf. Sci.*, 2020.
[34] W. Fan, S. Ma, N. Tang, and W. Yu. Interaction between record matching and data repairing. *ACM J. Data Inf. Qual.*, 2014.
[35] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *J. Amer. Statist. Assoc.*, 1969.
[36] M. Gebser, R. Kaminski, B. Kaufmann, and T. Schaub. *Answer Set Solving in Practice*. Morgan & Claypool Publishers, 2012.
[37] M. Gebser, B. Kaufmann, and T. Schaub. Multi-threaded ASP solving with clasp. *Theory Pract. Log. Program.*, 2012.
[38] P. Konda, S. Das, P. S. G. C., A. Doan, A. Ardalan, J. R. Ballard, H. Li, F. Panahi, H. Zhang, J. F. Naughton, S. Prasad, G. Krishnan, R. Deep, and V. Raghavendra. Magellan: Toward building entity matching management systems. *Proc. VLDB Endow.*, 2016.
[39] I. K. Koumarelas, T. Papenbrock, and F. Naumann. Mdedup: Duplicate detection with matching dependencies. *Proc. VLDB Endow.*, 2020.
[40] L. Li, J. Li, and H. Gao. Rule-based method for entity resolution. *IEEE Trans. Knowl. Data Eng.*, 2015.
[41] Y. Li, J. Li, Y. Suhara, A. Doan, and W. Tan. Deep entity matching with pre-trained language models. *Proc. VLDB Endow.*, 2020.
[42] Y. Li, J. Li, Y. Suhara, J. Wang, W. Hirota, and W. Tan. Deep entity matching: Challenges and opportunities. *ACM J. Data Inf. Qual.*, 2021.
[43] V. Lifschitz. *Answer Set Programming*. Springer, 2019.
[44] E. Livshits, A. Heidari, I. F. Ilyas, and B. Kimelfeld. Approximate denial constraints. *Proc. VLDB Endow.*, 2020.
[45] M. Manna, F. Ricca, and G. Terracina. Consistent query answering via ASP from different perspectives: Theory and practice. *Theory Pract. Log. Program.*, 2013.
[46] P. Nappa, D. Zhao, P. Subotic, and B. Scholz. Fast parallel equivalence relations in a datalog compiler. In *Proc. of PACT*, 2019.
[47] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records. *Science*, 1959.
[48] G. Papadakis, G. M. Mandilaras, L. Gagliardelli, G. Simonini, E. Thanos, G. Giannakopoulos, S. Bergamaschi, T. Palpanas, and M. Koubarakis. Three-dimensional entity resolution with JedAI. *Inf. Syst.*, 2020.
[49] E. H. M. Pena, F. Porto, and F. Naumann. Fast algorithms for denial constraint discovery. *Proc. VLDB Endow.*, 2022.
[50] S. Perri, F. Ricca, and M. Sirianni. Parallel instantiation of ASP programs: techniques and experiments. *Theory Pract. Log. Program.*, 2013.
[51] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. Linking data to ontologies. *Journal of Data Semantics*, 2008.
[52] A. Sahebolamri, L. Barrett, S. Moore, and K. K. Micinski. Bring your own data structures to datalog. *Proc. ACM Program. Lang.*, 2023.
[53] R. Singh, V. V. Meduri, A. K. Elmagarmid, S. Madden, P. Papotti, J. Quiané-Ruiz, A. Solar-Lezama, and N. Tang. Synthesizing entity matching rules by examples. *Proc. VLDB Endow.*, 2017.
[54] P. Singla and P. M. Domingos. Entity resolution with markov logic. In *Proc. of ICDM*, 2006.
[55] B. ten Cate, M. Funk, J. C. Jung, and C. Lutz. Fitting algorithms for conjunctive queries. *SIGMOD Rec.*, 2023.
[56] Z. Xiang, M. Bienvenu, G. Cima, V. Gutiérrez-Basulto, and Y. Ibáñez-García. ASPEN: ASP-based system for collective entity resolution. In *Proc. of KR*, 2024. Long version available at `https://arxiv.org/pdf/2408.06961`.
[57] Z. Xiang, M. Bienvenu, G. Cima, V. Gutiérrez-Basulto, and Y. Ibáñez-García. Advances in logic-based entity resolution: Enhancing ASPen with local merges and optimality criteria. In *Proc. of KR*, 2025. Long version available at `https:`

//github.com/zl-xiang/ASPEnP/blob/main/KR_2025_Supplementary_343.pdf.

[58] G. Xiao, D. Calvanese, R. Kontchakov, D. Lembo, A. Poggi, R. Rosati, and M. Zakharyaschev. Ontology-based data access: A survey. In *Proc. of IJCAI*, 2018.

Reminiscences on Influential Papers

This issue's contributors cover the impact of paying attention to the low-level implementation details, a paradigm shift in the way we approach stream processing, and the value of combining theoretical analysis with experimental evaluation. Furthermore, one of our contributors, rather than picking one paper, highlights the importance of putting the time to practice reading, reviewing, and learning from papers, not only from one's own field of interest but also from other fields. VLDB, similar to some systems conferences, launched a Shadow Program Committee for this purpose following the VLDB 2026 (Vol 19) submission cycles[1]. We wish to continue this effort in the future VLDB cycles. 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

Viktor Leis
Technical University of Munich
leis@in.tum.de

Stephen Tu, Wenting Zheng, Eddie Kohler, Barbara Liskov, and Samuel Madden.

Speedy Transactions in Multicore In-Memory Databases.

In Proceedings of the Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles (SOSP), 2013.

In 2013, as I was starting my PhD, the database community was in the middle of an in-memory DBMS wave. Projects like H-Store [12], HANA [6], and HyPer [7] were rethinking architecture for abundant main memory and many-core processors. H-Store championed a shared-nothing, partitioned design with single-threaded execution per partition – a beautifully simple approach when workloads partition cleanly. Silo [13] took a different path: a shared-everything single-node system in which any thread can access any data, a choice many considered inefficient and unscalable at the time.

The paper presents an in-memory transactional database optimized for modern multi-core CPUs. Its organizing principle is to minimize coordination among threads. Silo's key contribution is an OCC-based commit protocol that lets read-only transactions avoid shared-memory writes, relying on version checks rather than read latches for safety. Transactions perform writes locally and synchronize only at commit. To avoid deadlocks, Silo sorts the write set and acquires locks in that order; conflicts simply trigger aborts.

What stood out to me was the care in low-level implementation. Each record carries a 64-bit TID word that encodes the commit timestamp plus lock / status bits, enabling a thread to lock and update a version in one atomic operation. Silo also addresses phantoms without heavy locking: leveraging Masstree's [9] node versioning, a transaction records the nodes it scanned and, at commit, aborts if any of those node versions changed – preventing range anomalies while still avoiding next-key locks.

Silo scaled extraordinarily well on a 32-core machine, reaching 700K TPC-C transactions per second – much higher throughput than prior reports at the time. The experiments were unusually thorough, including ablations that explain why it performs well, and the open-source code allowed others to learn from and build upon their implementation.

The paper strongly influenced my approach to

[1] https://vldb.org/2026/shadow-pc.html

systems research. It taught me that co-designing DBMS components often considered in isolation – such as concurrency control, latching, and indexing – can yield dramatic performance gains. It also showed me that on modern hardware, seemingly small low-level implementation choices can have outsized performance implications.

Anja Gruenheid
Microsoft Gray Systems Lab, Switzerland
anja.gruenheid@microsoft.com

In this column, I do not want to single out one paper but rather share a few experiences that helped me grow as a researcher by reading classic, influential works as well as learning to engage with and review ongoing research from others. During my Ph.D., I was fortunate to be part of activities that encouraged us to step outside our narrow focus and explore ideas from different corners of systems research. Looking back, those moments, reading papers from areas intersecting with databases, discussing them openly, and later practicing reviewing in a low-stakes setting, taught me lessons no formal course ever could. They showed me that many skills we often take for granted as researchers, like forming opinions about a paper and giving constructive feedback, do not simply emerge on their own. They need space, practice, and above all, a mindset that values understanding over judgment. As I started engaging more with research beyond my immediate area, I realized how much we can learn simply by looking closely at work that is different from our own. It is easy to assume that writing a paper equips us to judge another, but in reality, the real benefit of reading broadly is not about evaluation alone, it is about perspective. Systems research is wonderfully diverse, as databases intersect with distributed systems, networking, hardware, and now machine learning. Every paper reflects choices shaped by these contexts, and appreciating those choices requires us to step outside our familiar ground. This felt challenging early in my career as parts of a paper would sometimes seem only partly within reach. Over time, though, I came to see these moments not as obstacles but as opportunities to expand my understanding of the field. That broader view, in turn, makes both research and reviewing richer and more thoughtful. These are not skills that appear automatically, they develop through deliberate practice and above all, a willingness to approach unfamiliar ideas with curiosity.

That realization brings me to the experiences that shape such habits of mind. Early in my Ph.D., I had the good fortune to be part of an effort that explicitly allowed us students to learn from classic research papers and sharpen our reviewing skills at the same time. The professors in my group recognized that the ability to review well stems from perspective, an understanding not only of technical details but of the broader landscape and its history. To that end, they curated a list of about twenty papers spanning databases, operating systems, networking, and distributed systems, works that had left a lasting imprint on the field. I remember learning how statistics like histograms and sampling play a vital role in query plan generation and indexing strategies through the work of Chaudhuri and colleagues, which provided fascinating insights into how DBMS actually leverage these techniques. I also encountered Lamport's work on Paxos, a paper whose ideas, to my surprise, I would see surface in different guises again and again as a reviewer. And then there were papers on topics such as microkernels, which I have not really crossed paths with since, yet they opened my eyes to the rigor and elegance of fundamental research in adjacent communities. Despite several attempts, I have not been able to locate the original list. In hindsight, though, the specific papers mattered less than the way we engaged with them. For each paper, we organized a discussion session. A graduate student, chosen at random, would present the work, explain its contributions, and lead the conversation. We would ask what problem the paper was trying to solve, why that problem mattered in its original context, what conceptual leap the authors had made, and to what extent their ideas influenced the systems that came after. The element of randomness was important, as it meant that any of us could be the presenter for any given paper, which in turn meant that all of us needed to come well prepared. Skimming was not an option. At the time, this felt demanding. After all, each of us had our own work, deadlines to chase, and code to debug. Spending hours poring over a decades-old paper on distributed operating systems when your own work was on data integration might seem like an indulgence. But this was no indulgence, it was, in retrospect, a gift. Those sessions pushed us beyond the confines of our chosen topics. They taught us intellectual humility and curiosity, a desire to appreciate the reasoning behind design decisions, algorithmic choices, and experiments crafted under very different assumptions about hardware and software than those that dominate today.

Looking back, I see several reasons why this experience mattered so much. It instilled habits of critical thought, yes, but also modeled a tone for that critique, one that aimed to understand rather than dismiss. And perhaps without our noticing at the time, it established a standard for clarity, as we began to see that the most influential papers were often those that told their story with simplicity and precision even when the underlying idea was subtle or complex. Above all, it showed us a truth that extends beyond any one field, that breadth is not the enemy of depth but its complement, and that the discipline of engaging seriously with ideas outside one's immediate path strengthens one's ability to make meaningful contributions within it. Of course, not every group can replicate this exact format, and not every advisor has the time to lead such sessions regularly. But if we agree that good reviewing matters and by extension, that the quality of dialogue in our conferences and journals matters, then we as a community must think creatively about how to offer similar opportunities. The responsibility does not rest with advisors alone. We all have a shared stake in this process because the benefits ripple outward, strong reviewers make for stronger feedback, which makes for stronger papers, which makes for a stronger field. Finding practical ways to create such learning experiences is not always straightforward but it is possible.

One mechanism I have come to value deeply in recent years is the shadow program committee. For those unfamiliar, a shadow PC runs in parallel with the official review process of a conference. Its members read and review the same papers, often following the same guidelines, and later compare their assessments with the real decisions. When I was a student, I joined a EuroSys shadow PC and found the experience transformative. Until then, I had thought of reviewing as a mostly solitary act, you read, you form an opinion, you write it down. What I saw instead was the collective effort that underlies every acceptance and rejection. I saw reviewers with different backgrounds weigh novelty in different ways. I saw discussions wrestle with incomplete evaluations, ambiguous claims, or competing intuitions about practicality versus elegance. And I saw how much thought goes into offering feedback that is both candid and constructive. In addition to EuroSys, we also organized an internal shadow PC in our group for SoCC papers to practice good reviewing practices, and that too left an enduring impression. Together, these experiences made me not only a better reviewer but also a better author. Understanding the questions reviewers routinely ask, such as What gap does this work fill? What prior work does it build upon or overlook? How do the experiments support the claims? helped me anticipate and address them in my own submissions. Shadow PCs offer a rare opportunity, they involve students in real decisions without real stakes and demystify a process that can otherwise feel opaque. In doing so, they reinforce a message we should all embrace, reviewing is a craft, and like any craft, it can be taught, practiced, and improved.

What stayed with me most from those early exercises was not just the specific ideas in any single paper but the habit of grappling with work outside my immediate comfort zone. Reading and truly trying to understand papers that fell well beyond the oftentimes narrow boundaries of my research opened my eyes to the sheer diversity of questions and approaches that systems research embraces. It taught me that there is no single mold for what constitutes important or elegant work, different subfields prize different virtues, and appreciating those differences deepens both our perspective as researchers and our fairness as reviewers.

Paris Carbone

KTH Royal Institute of Technology & RISE Research Institutes of Sweden, Sweden

parisc@kth.se

Tyler Akidau, Robert Bradshaw, Craig Chambers, Slava Chernyak, Rafael J. Fernández-Moctezuma, Reuven Lax, Sam McVeety, Daniel Mills, Frances Perry, Eric Schmidt, and Sam Whittle.

The Dataflow Model: A Practical Approach to Balancing Correctness, Latency, and Cost in Massive-Scale, Unbounded, Out-of-Order Data Processing.

In Proceedings of the VLDB Endowment, 2015.

In the summer of 2015, beneath the tropical skies of the Kohala Coast, the conference room at Hilton Waikoloa Village was packed with anticipation. There, Tyler Akidau and fellow engineers took the stage, unveiling Google's ambitious quest to harmonize the realms of batch and streaming. VLDB 2015 was my very first exposure to the data management community as a fresh graduate student. Among a packed schedule that summer, three sessions stood out: Michael Stonebraker's journey at his turing award seminar, Peter Bailis' inspiring talk of coordination avoidance, and most captivating of all,

Google's kaleidoscopic presentation [2] of "The Dataflow Model" [2]. This column reflects on why the latter mattered so much then and why, even after ten years, it continues to shape our thinking.

What Tyler and colleagues called for, was a much-needed paradigm shift: *streaming should subsume batch*. Their model reads more like a manifesto for out-of-order processing; every element is stamped with its event-time, and windowing can be applied uniformly across both bounded and unbounded inputs. A bounded dataset is thus treated as a special case of an unbounded one. Dataflow discourages the runtime-specific terms "streaming" and "batch" in favor of the more precise "unbounded" and "bounded" datasets. The model evolved from two existing Google systems: FlumeJava [4] and MillWheel [1]: their lower-abstraction level dataflow runtime. Dataflow featured a somewhat overly lean unified programming abstraction based on two primitives: ParDo for parallel processing functions and GroupByKey for keyed grouping. These behave identically in both batch and streaming settings. The subsequent release of Apache Beam as the open source incarnation of Dataflow reinforced the sense that this was the product of a determined engineering team rather than a traditional industrial research group.

When the first bold statements by Tyler hit the stage the reactions were polarized. One could tell that from the diverse facial expressions of the audience. The Q&A session confirmed some of my own initial concerns. To some, naming the model "Dataflow" felt like appropriation. Dataflows indeed have a long history. Others highlighted that the model was restrictive and offered nothing fundamentally new. From a purist point of view the dataflow model deliberately ignored a wide set of complex data stream window types researchers have been building towards for decades. Yet, to younger me the simplicity was a revelation. I had just spent my first PhD year wrestling with ad-hoc window semantics, experimenting with every imaginable combination of complex windows. Google's model explicitly limited this proliferation: fixed, sliding and session windows were the canonical choices. Besides, "time" was the only dimension that mattered for correctness, beautifully captured using watermarks, triggers and accumulation modes. Personally, I remember feeling equal parts relieved and irritated: relieved to see a coherent framework for reasoning about time and correctness, and mildly irritated that much of the prior work on richer window types seemed destined to take a back seat.

Where the Dataflow model positioned itself was at the intersection of long-standing database theories and the pragmatic demands of cloud-scale applications. It was not a matter of industry versus research, but of industry distilling a favored set of research ideas into products that would endure. Much of the "Dataflow Model" drew on the influential contributions of David Mayer et al. [8] in stream processing. The model also built on the sophisticated MillWheel/Dataflow runtime [1], which delivered unprecedented performance for transactional, stateful streaming workloads. This combination left little room for competition; convergence, it seemed, was inevitable and just over the horizon.

How these ideas reshaped systems and research became evident in the evolution of Apache Flink, Kafka Streams, Spark Streaming, and their peers. Several experimental Flink window types that fellow committers and I had added only months earlier had to be rewritten or removed to conform to the deterministic semantics championed by Dataflow. In retrospect, this process resembled what Schumpeter described as *creative destruction* in "Capitalism, Socialism and Democracy" [11]; tearing down existing designs was painful at the time, yet it cleared the way for a stronger and more coherent foundation for stream processing. The shift spread through research and industry like a major wave of innovation: Ververica (then Data Artisans) launched its "out-of-order" alignment mission, and shortly after Databricks and Confluent adopted similar principles in Structured Streaming and Kafka Streams respectively. Within just a year, the community's vocabulary and expectations had converged on event time, watermarks, and bounded versus unbounded data as the universal frame of reference. This was a truly impressive impact feat on its own. Apache Flink emerged as a popular runner of Apache Beam, Google's open incarnation of the Dataflow model, and this alignment greatly accelerated its industry adoption. At the same time, because the Dataflow model did not prescribe how a runtime should operate internally, much of the foundational work we had done on Flink, such as state checkpointing, remained not only relevant but essential, and continues to be so today [3, 10].

A decade later, the waters remain calm, perhaps too calm. No shift since has reshaped cloud data processing as profoundly as the Dataflow model. As for the runtimes, disaggregated state is now making a comeback with Flink 2.0 [10], an architecture el-

[2]The introductory slide deck for Google Dataflow is often remembered as a communication marvel in its own right: dark-mode, flashy, with precise animations that built up complex data processing ideas in a simple, understandable, and slightly "trippy" way.

ement already present in the very first version of Millwheel.

Eleni Tzirita Zacharatou
Hasso Plattner Institute & University of Potsdam, Germany
eleni.tziritazacharatou@hpi.de

Chee-Yong Chan and Yannis E. Ioannidis.
Bitmap Index Design and Evaluation.
In Proceedings of the International Conference on Management of Data (SIGMOD), 1998.

I first encountered this paper [5] in 2013 during the exploratory phase of my PhD, when I was searching for a research direction as a newcomer to the database field. Although I ultimately did not work much in the area of bitmap indexing, this study of bitmap indexes left a lasting impression during those formative first months of my PhD journey as I began to understand the landscape of database research. Beyond its significant technical contributions, Chan and Ioannidis's paper served as a model for how I approach problems, structure my research methodology, and communicate my findings. Essentially, their work was instrumental in shaping my understanding of what constitutes high-quality database research.

In today's research taxonomy, Chan and Ioannidis's paper would be classified as an Experiments and Analysis (E&A) study, but one that goes well beyond conventional experimental evaluation by incorporating both analytical modeling and novel algorithmic contributions. The work presents a comprehensive framework for understanding the design space of bitmap indexes, systematically investigating key design dimensions including attribute value decomposition and encoding approaches, selection query algorithms, and compression and caching techniques. Their analysis identified four critical points in the space-time tradeoff curve – ranging from space-optimal to time-optimal configurations – and provided what they described as "a first set of guidelines for physical database design using bitmap indexes." In subsequent years, bitmap indexes became widely adopted in commercial systems like Oracle for data warehousing, were a core component in early column stores, and powered specialized libraries such as FastBit.

What stood out to me was the paper's effective combination of theoretical analysis and practical evaluation. This inspired me early in my research career to always strive for principled analysis alongside thorough experimental validation. But more fundamentally, this paper taught me an important research philosophy: the value of taking a step back before moving forward. Chan and Ioannidis demonstrated that truly understanding the current landscape, identifying existing trade-offs, and systematically mapping the design space are essential tools for guiding innovation. The elegance of their framework lies not just in organizing existing knowledge, but in articulating the underlying design principles in a way that reveals previously unconsidered alternatives.

Perhaps most profoundly, the paper illustrated the power of abstraction and decomposition in research. By breaking bitmap indexing into its fundamental elements, the authors enabled new compositions and revealed hidden trade-offs. This taught me that understanding complex systems requires first decoupling their components, and that this decoupling process itself often illuminates the path forward.

The combination of theoretical analysis and practical evaluation in the paper is evident not only in its content but also in its structure. Rather than grouping all experiments in a separate section, as is standard practice today, Chan and Ioannidis interleave experimental validation with analytical insights throughout the paper. This structure creates a more coherent reading experience, as each theoretical result is immediately supported by relevant experimental evidence, allowing readers to follow the argument without having to switch between different sections.

In conclusion, I believe this paper demonstrates that a careful analysis of trade-offs in physical database design is essential to database research. For anyone looking to understand bitmap indexing, or more importantly, to learn how to conduct a systematic design space exploration, this paper is an excellent guide.

REFERENCES

[1] Tyler Akidau, Alex Balikov, Kaya Bekiroğlu, Slava Chernyak, Josh Haberman, Reuven Lax, Sam McVeety, Daniel Mills, Paul Nordstrom, and Sam Whittle. MillWheel: Fault-Tolerant Stream Processing at Internet Scale. *Proc. VLDB Endow.*, 6(11):1033–1044, August 2013.

[2] Tyler Akidau, Robert Bradshaw, Craig Chambers, Slava Chernyak, Rafael J. Fernández-Moctezuma, Reuven Lax, Sam McVeety, Daniel Mills, Frances Perry, Eric

Schmidt, and Sam Whittle. The Dataflow Model: A Practical Approach to Balancing Correctness, Latency, and Cost in Massive-Scale, Unbounded, Out-of-Order Data Processing. *Proc. VLDB Endow.*, 8(12):1792–1803, August 2015.

[3] Paris Carbone, Stephan Ewen, Gyula Fóra, Seif Haridi, Stefan Richter, and Kostas Tzoumas. State Management in Apache Flink®: Consistent Stateful Distributed Stream Processing. *Proc. VLDB Endow.*, 10(12):1718–1729, August 2017.

[4] Craig Chambers, Ashish Raniwala, Frances Perry, Stephen Adams, Robert R. Henry, Robert Bradshaw, and Nathan Weizenbaum. FlumeJava: Easy, Efficient Data-Parallel Pipelines. In *Proceedings of the 31st ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI '10, page 363–375, New York, NY, USA, 2010. Association for Computing Machinery.

[5] Chee-Yong Chan and Yannis E. Ioannidis. Bitmap Index Design and Evaluation. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, SIGMOD '98, page 355–366, New York, NY, USA, 1998. Association for Computing Machinery.

[6] Franz Färber, Sang Kyun Cha, Jürgen Primsch, Christof Bornhövd, Stefan Sigg, and Wolfgang Lehner. SAP HANA Database: Data Management for Modern Business Applications. *SIGMOD Rec.*, 40(4):45–51, January 2012.

[7] Alfons Kemper and Thomas Neumann. HyPer: A Hybrid OLTP&OLAP Main Memory Database System Based on Virtual Memory Snapshots. In *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, ICDE '11, page 195–206, USA, 2011. IEEE Computer Society.

[8] Jin Li, Kristin Tufte, Vladislav Shkapenyuk, Vassilis Papadimos, Theodore Johnson, and David Maier. Out-of-Order Processing: A New Architecture for High-Performance Stream Systems. *Proc. VLDB Endow.*, 1(1):274–288, August 2008.

[9] Yandong Mao, Eddie Kohler, and Robert Tappan Morris. Cache Craftiness for Fast Multicore Key-Value Storage. In *Proceedings of the 7th ACM European Conference on Computer Systems*, EuroSys '12, page 183–196, New York, NY, USA, 2012. Association for Computing Machinery.

[10] Yuan Mei, Rui Xia, Zhaoqian Lan, Kaitian Hu, Lei Huang, Paris Carbone, Yanfei Lei, Vasiliki Kalavri, Han Yin, and Feng Wang. Disaggregated State Management in Apache Flink® 2.0. *Proc. VLDB Endow.*, 18(12):4846–4859, September 2025.

[11] Joseph A. Schumpeter. *Capitalism, Socialism and Democracy*. Routledge: London, UK, 1976.

[12] Michael Stonebraker, Samuel Madden, Daniel J. Abadi, Stavros Harizopoulos, Nabil Hachem, and Pat Helland. The End of an Architectural Era: (It's Time for a Complete Rewrite). In *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB '07, page 1150–1160. VLDB Endowment, 2007.

[13] Stephen Tu, Wenting Zheng, Eddie Kohler, Barbara Liskov, and Samuel Madden. Speedy Transactions in Multicore In-Memory Databases. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, SOSP '13, page 18–32, New York, NY, USA, 2013. Association for Computing Machinery.

ADVICE TO MID-CAREER RESEARCHERS

Selected Statements on the Academic Enterprise

Christian S. Jensen, Aalborg University, Denmark

I start with a disclaimer: I do not think it is my business to tell others what to do. Rather, how we choose to think and act are our personal responsibility. This said, I am happy sharing my observations and views and, in that sense, giving advice. Others may consider this as input when deciding on how to think and act.

In the following, I will comment on a variety of topics that are relevant to our practices as academics. When preparing this document, I first noted down a list of candidate topics. Then I eventually chose a selection of those for inclusion.

Be nice. I recommend being nice to others or at least trying to. We all have different backgrounds and are formed by different experiences, and what does not make sense or seem reasonable to one person may make sense or look reasonable to others.

Let me give an example to illustrate this statement. You might have attended a talk where a young scientist presents research that you think is problematic. The research may make inappropriate assumptions, it may make claims that are not substantiated well, it may ignore some related work, or it may simply be presented poorly. In such a situation, it is easy to get offended and to hang the presenter out to dry! But why? The person has likely tried their best and may think that the research and presentation are fine or at least the best possible, given the circumstances. In this situation, it is best to politely ask the presenter whether, e.g., it is possible to clarify specific assumptions or to comment on the relation to another line of research. This way, it is possible to flag to the presenter or knowledgeable participants in the audience that something perhaps needs to be looked further into, and the presenter is given the opportunity to argue for their research and clarify any misunderstandings. It is also possible to talk with the presenter after the presentation. This can all be done in a supportive and constructive manner.

A key point to realize is that one can be nice without lowering one's standards. Often, when one has something critical to say, it is best said one-to-one. And if you can include positive remarks as well, the person you criticize is much more likely to listen, and you have been effective and have not wasted your time. There are times when it is best to simply move on and leave it to others. Choose your battles carefully.

Overall, being nice is good for the community as well as for oneself.

Understand that research is a social activity. There are many aspects to this observation. Growing up as a scientist, I benefitted tremendously from being part of a community, including getting new ideas, insights, directions; being able to form collaborations; and obtaining letters of recommendation. Being located at a small university in a small country, I realized early on that I had to engage in community efforts. For example, I served on many program committees and in a variety of other roles at conferences and beyond. I also attended both the top conferences in my general area and specialized conferences that aligned with my specific research focus. I recommend that you find a community and then invest in being part of it.

Another aspect is that the world is surprisingly small. People you meet once, you will often meet again, even if you did not think so at the time. This is yet a reason for being nice.

At the smaller scale of specific research collaborations, our research is also a social activity. Certainly, my collaborators keep me going…and keep me very busy. At this level, it is important to be responsible and supportive. This way, your group of collaborators will grow. So, it is not good to frequently be missing in action – busy with something else – when the real work needs to be done. Collaborators see through that and eventually move on. This leads to the next topic.

Say no, sometimes. I do not know about you, but I sometimes find it hard to say no. But I am at least getting better at it. It is hard to say no when presented with a

concrete opportunity that one finds meaningful, and when it will be months into the future before something must be done. But, of course, choosing to do something means that there is something else that one cannot do, either work-related or outside of work. Yet, that "something else" is vague, and the calendar looks relatively open months from now. And saying yes will be good for your career. It is easy to say yes – the hard part of delivering only comes later.

I was talking with a colleague about this recently. The colleague made the point that one should ask oneself: Would I still say yes if I had to do the work this or next week, rather than some months from now? If the answer is no, say no. As I agree that it is an illusion that we will somehow have an open schedule some months from now, this is a very good point. Sometimes, saying yes too often can even jeopardize one's ability to deliver on what one has already said yes to.

This brings me to the issue of providing service to the scientific community. We should all provide such service. Given this, it is best to provide service where it matters the most. This is often where the quality standards are the highest. An important part of service is to be part of program committees and to review for journals. I like to distinguish between four categories of reviewers: (i) those who say no, (ii) those who say yes and do the work in a timely fashion, (iii) those who do not deliver on time, but eventually do deliver, and (iv) those who disappear or keep saying that they will deliver but never do. Since reviewing is volunteer work and since we are all busy, there should be some flexibility. But being often in category iii and, certainly, category iv is not good for anybody. Reviewers in these categories cause unnecessary problems, and the reviewers risk getting a bad reputation. Why spend the time and hurt your reputation in the process? It does not take more time to do timely reviewing.

Balance continuity and renewal. It is an important consideration to put effort into finding and maintaining a productive balance between continuity and renewal in one's research. The right balance surely varies from person to person, and sometimes one needs to go with the flow. Transitioning too infrequently can render one's research uninteresting, and transitioning too frequently can compromise quality and depth.

I started out working on temporal databases, and this line of research remained my focus for a decade. Then I got involved in a project on spatiotemporal databases. This led to work on the indexing of spatiotemporal data, where I was able to build on what I had learned from working on the indexing of temporal data. We also started to see the contours of the mobile revolution that led to roughly everybody having a mobile phone. Thus, I transitioned to working primarily on data management and query processing for what we called "moving objects." Later, motivated by the proliferation of geo-textual content, spatial keyword querying became a primary activity. This was subsequently replaced by work on the use of spatial trajectories, which continue to proliferate, for a variety of purposes, including vehicle routing. The latest main activity, motivated in large part by the growing Internet of Things and the deployment of sensors throughout industry and society, is time-series analytics, where neural technologies play a key role.

Each time I made a transition, I was able to build on what I had learned from my previous research. And the transitions often occurred because of, or as part of, collaborations with colleagues.

Find unexplored territories. The life of a researcher working in an overpopulated area is a difficult one. Towards the end of when I worked primarily on temporal databases, the literature contained numerous proposals for temporal data models and query languages. Proposing a new one was an uphill battle. One needed to compare to many existing proposals, each with at least one very strong proponent. It was increasingly difficult to do something substantially different and better, let alone convince reviewers of this.

The life of a researcher working in an unexplored territory is comparatively easier. One does not need to implement and compare with a proliferation of existing proposals, and the prospects for performing novel and impactful research are much better.

When I worked in temporal databases, we were dealing with two temporal aspects of data: when the data was true in reality and when the data was recorded as current in the database. Such data could be true from some time in the past until the current time, *now*. Likewise, data was part of the current database state from when it was

inserted until it was deleted or updated. These temporal aspects could be viewed as two-dimensional regions that grew continuously over time. We had worked on the indexing of such data and then saw, as mentioned already, the contours of a mobile Internet of users capable of continuous movements. This led to the question of how we could index moving objects. This, in turn, led us, and other members of the community, to a territory where objects could move continuously rather than being stationary. Here, we needed new solutions for indexing and query processing, e.g., for range and nearest neighbor queries. It was indeed a new territory full of new challenges. For starters, everything that had been done for static points, we could consider doing for moving objects. It was an exciting time.

Later, combining text with spatial data, including moving objects, again opened a new territory, as did the use of trajectory data in transportation and other urban applications, including for routing, where data from fixed, in-road sensors was previously the primary or only data source. Finally, with tens of zettabytes of streaming data being generated annually by IoT devices, there are unmet challenges to value creation from time-series data at scale.

Seek flow. When I was younger, I worked late and got up late when possible. My rule was that I should get to the office no later than noon. I would work until dinner. After dinner, I would go back to the office and work until, say, 3 a.m. Early in my career, I spent four sabbaticals with Rick Snodgrass at the University of Arizona, and I have fond memories of the many late nights working in the lab. I liked Led Zeppelin (I still do), and I remember putting on a CD (yes, we had CDs) to listen to specific songs. Then I would continue working, only to realize at some point that no music was playing without having any memory of having listened to the songs I wanted to listen to. When working those nights, time and everything else often disappeared, and only the work was in focus. I have had the same kind of experience before and after these sabbaticals. I found, and still find, this to be very relaxing, almost therapeutic. I later learned that this phenomenon is called flow and has been studied extensively by psychologists, although I have yet to read about it. Still, I recommend trying to find flow.

Oh, I still end up working late, although I do it from home. This is often because I have said yes to too much and because of the social nature of conducting research!

Navigating the Performance-Security Trade-Off in Future Analytics on Shared Data

Zsolt István
Systems Group, TU Darmstadt, Germany
zsolt.istvan@tu-darmstadt.de

Securing analytics on shared data is important but expensive. Analyzing datasets from multiple data owners can yield valuable insights [1, 2, 3, 4, 5] but poses significant security risks. Even within enterprises – our primary focus – precautions are necessary when handling data across subsidiaries and geographic regions [6, 7]. Existing security solutions based on Trusted Execution Environments (TEEs) [8, 9], fully homomorphic encryption [10], and structured encryption [11] offer strong protections, albeit in a physically centralized manner. For more decentralization, there are exciting approaches based on Secure Multi-Party Computation (MPC) [12] that do not need a trusted third party nor merging datasets at a central location. Recent projects [6, 13, 14, 15] show that MPC can reduce the risk of leaks for analytics on shared data under stronger security guarantees. However, MPC queries are often impractically slow, requiring orders of magnitude more computation and communication than plain-text or TEE-based query execution.

Adding security measures is a balancing act in the enterprise. Conventional wisdom dictates not to compromise on security between distrusting parties at all – no matter the performance impact. In the context of in-house analytics at large enterprises, however, even if only parts of a query are run with improved security, there is already a benefit for the enterprise [6]. Adding protection through the use of TEEs and MPC to the existing DBMS-level ones is useful if performance does not plummet, and future databases should be able to decide, given a performance target, what level of security can be actually fulfilled.

Analytics on shared data need security-aware query planning. We are working on a platform that modularizes secure query execution and allows for different strategies for trading off performance and security at the operator and query level. One point in the trade-off space is protecting computation using TEEs: we are exploring how to run OLAP queries in TEEs without performance overhead [8]. Another solution is using MPC and we are investigating how to precisely control information leakage about data passing between operators in exchange for faster MPC query execution. In the future, the query planner will need to be able to combine local and distributed operators executing in plain-text, in TEEs, using MPC, etc., and under different adversarial models. For completeness, in addition to the systems-level challenges, it will be also necessary to define security levels that are tailored to DBMS use-cases.

Case study: Trading off intermediate result size protection for better performance. As a concrete example of trading off security for performance, consider how intermediate results are passed between operators in an MPC query. The execution of MPC operators is oblivious to the content of their input: an oblivious filter, for instance, produces an output equal in size to its input but with a secret column indicating which row is actually selected. Similarly, an oblivious join has an output size equal to the Cartesian product of its inputs. This results in data sizes snowballing as the query execution proceeds, especially for analytical queries with many joins [6, 14, 15, 16]. Related work explores the relaxation of intermediate result size protection in different ways, e.g., adding non-deterministic noise to the true intermediate result size [14] or entirely foregoing adding noise to it [6]. One common decision, however, is to combine the implementation of the intermediate result size protection with the actual operator logic.

In Reflex [17] we decouple the protection mechanism from the operator logic, achieving flexibility while retaining execution efficiency thanks to a highly parallel implementation. The benefit of implementing intermediate result size protection as a separate step after each oblivious operator is that we can define custom strategies for hiding the size of the intermediate results and, through this, offer different security/performance trade-offs. These strategies could be based on related work, using, e.g., differentially private noise [14], or entirely new ones. Reflex approaches secure shared analytics differently from most related work: instead of prescribing a specific set of security guarantees, we build the mechanisms necessary for the query planner to pick the adequate protections for each query based on performance and security/privacy criteria.

REFERENCES

[1] D. W. Archer, D. Bogdanov, Y. Lindell, L. Kamm, K. Nielsen, J. I. Pagter, N. P. Smart, and R. N. Wright, "From keys to databases—real-world applications of secure multi-party computation," *The Computer Journal*, vol. 61, no. 12, pp. 1749–1771, 2018.

[2] D. Bogdanov, L. Kamm, B. Kubo, R. Rebane, V. Sokk, and R. Talviste, "Students and taxes: a privacy-preserving study using secure computation," *Proceedings on Privacy Enhancing Technologies*, 2016.

[3] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

[4] A. Rajan, L. Qin, D. W. Archer, D. Boneh, T. Lepoint, and M. Varia, "Callisto: A cryptographic approach to detecting serial perpetrators of sexual misconduct," in *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '18, (New York, NY, USA), Association for Computing Machinery, 2018.

[5] N. Crooks, "Efficient data sharing across trust domains," *ACM SIGMOD Record*, vol. 52, no. 2, pp. 36–37, 2023.

[6] W. Fang, S. Cao, G. Hua, J. Ma, Y. Yu, Q. Huang, J. Feng, J. Tan, X. Zan, P. Duan, Y. Yang, L. Wang, K. Zhang, and L. Wang, "Secretflow-scql: A secure collaborative query platform," *Proc. VLDB Endow.*, vol. 17, p. 3987–4000, Nov. 2024.

[7] S. Becker, C. Bösch, B. Hettwer, T. Hoeren, M. Rombach, S. Trieflinger, and H. Yalame, "Multi-party computation in corporate data processing: Legal and technical insights." Cryptology ePrint Archive, Paper 2025/463, 2025.

[8] A. Lutsch, M. El-Hindi, M. Heinrich, D. Ritter, Z. István, and C. Binnig, "Benchmarking analytical query processing in intel sgxv2," in *Proceedings 28th International Conference on Extending Database Technology, EDBT 2025, Barcelona, Spain, March 25-28, 2025* (A. Simitsis, B. Kemme, A. Queralt, O. Romero, and P. Jovanovic, eds.), pp. 516–528, OpenProceedings.org, 2025.

[9] P. Antonopoulos, A. Arasu, K. D. Singh, K. Eguro, N. Gupta, R. Jain, R. Kaushik, H. Kodavalla, D. Kossmann, N. Ogg, *et al.*, "Azure sql database always encrypted," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 1511–1525, 2020.

[10] S. Bian, Z. Zhang, H. Pan, R. Mao, Z. Zhao, Y. Jin, and Z. Guan, "He3db: An efficient and elastic encrypted database via arithmetic-and-logic fully homomorphic encryption," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2930–2944, 2023.

[11] R. A. Popa, C. M. Redfield, N. Zeldovich, and H. Balakrishnan, "Cryptdb: Protecting confidentiality with encrypted query processing," in *Proceedings of the twenty-third ACM symposium on operating systems principles*, pp. 85–100, 2011.

[12] R. Cramer, I. B. Damgård, *et al.*, *Secure multiparty computation*. Cambridge University Press, 2015.

[13] J. Liagouris, V. Kalavri, M. Faisal, and M. Varia, "SECRECY: Secure collaborative analytics in untrusted clouds," in *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pp. 1031–1056, 2023.

[14] J. Bater, X. He, W. Ehrich, A. Machanavajjhala, and J. Rogers, "Shrinkwrap: efficient sql query processing in differentially private data federations," *Proceedings of the VLDB Endowment*, vol. 12, no. 3, 2018.

[15] J. Bater, Y. Park, X. He, X. Wang, and J. Rogers, "Saqe: practical privacy-preserving approximate query processing for data federations," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 2691–2705, 2020.

[16] Y. Zhang, J. Bater, K. Nayak, and A. Machanavajjhala, "Longshot: Indexing growing databases using mpc and differential privacy," *Proceedings of the VLDB Endowment*, vol. 16, no. 8, pp. 2005–2018, 2023.

[17] L. Gu, S. Zeitouni, C. Binnig, and Z. István, "Reflex: Speeding up smpc query execution through efficient and flexible intermediate result size trimming." `https://arxiv.org/abs/2503.20932`, 2025.

Sihem Amer-Yahia Speaks Out on Social Computing and DEI

H. V. Jagadish and Vanessa Braganholo

Sihem Amer-Yahia

https://lig-membres.imag.fr/amery/

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. Sihem Amer-Yahia is my guest today. She is a Silver Medal Research Director at the French National Center for Scientific Research (CNRS) and Deputy Director of the Laboratoire d'Informatique de Grenoble, one of the largest research labs in Computer Science in France, with CNRS and INRIA Researchers and University Professors. She has won many awards, including the 2024 IEEE TCDE Impact Award, the ACM SIGMOD Contributions Award, and the VLDB Women in Database Research Award. Welcome, Sihem!

Thank you, Jag, for the invitation and for the introduction. And let me also thank you for giving me a chance in your team as a postdoc at AT&T Labs in 1999. That was the start of the first chapter of my work life.

You have made significant contributions in many areas, but if I had to pick one, I would probably name social computing in the context of data management. Can you tell us a little about how you came to this topic and what work you're most proud of in this area?

In my career, I went from core data management questions, such as query processing for structured and unstructured data, to exploiting subjective human-generated data. My transition to social computing started when I joined Yahoo! Research in 2006. It was a time when the Web 2.0 was burgeoning. Web application owners understood the need for social interactions to drive traffic to their site, and pure social network developers understood the value of content. I quickly understood the importance of the social nature of data produced by humans, and I became convinced that to build applications where humans interact effectively with each other and with data, it was necessary to think of data models that capture human factors and behavior. In 2009, I wrote a CIDR paper with Cong Yu and Laks Lakshmanan titled "SocialScope: Enabling Information Discovery on Social Content Sites," where we introduced a graph data model, an algebra to manipulate data about people and their interactions, and a system architecture to build applications on the social Web.

The amount of effort our community has deployed for all sorts of questions around XML storage, query design, processing algorithms, optimization, and later on, XML Full Text Search is tremendous. Even though that did not "make it" in the same way relational databases made it, I believe it had a wider and lasting impact on us as a research community.

At that time, the work that was going on in social computing was very much observational. I went to conferences such as CSCW and ICWSM, and I met social scientists and psychologists. I became aware of the fact that there were so many theories in the social sciences that could be verified on the social Web. The hard question was which of those theories mattered and how to distil them into questions that mattered to me as a database researcher.

The work I'm most proud of in social computing is relatively recent because some of those theories only started to make it into my own research 10 years later, in 2017. It happened in the work of my student, Julien Pilourdault, where we examined the impact of human factors on designing recommendation algorithms on online crowdsourcing and labor markets. We read a lot about theories from the Psychology of Work that date back to the 70's, where they solved everything about people at work and their motivation, and we used that to formalize intrinsic and extrinsic motivation factors. That helped us better understand how to design adaptive algorithms that observe people as they complete tasks and capture their motivation to feed it into the logic of recommendations. That work required a fair amount of engineering, too, and led to a collaboration with Atsuyuki Morishima at the University of Tsukuba to build Crowd4U, an academic crowdsourcing platform. In retrospect, we only addressed the tip of the iceberg.

Well, that's a lot, though!

You did some really impactful work on XML early in your career, including major contributions to XPath and a highly cited paper on tree patterns. At that time, we all thought XML was going to take over the universe. That hasn't quite happened, though there is a very solid niche for XML today. What is your opinion about that?

The amount of effort our community has deployed for all sorts of questions around XML storage, query design, processing algorithms, optimization, and later on, XML Full Text Search is tremendous. Even though that did not "make it" in the same way relational databases made it, I believe it had a wider and lasting impact on us as a research community. The work of XML in the database (DB) community initiated a movement in our community. It made us rethink the fundamentals of query processing. It made us relevant to the Information Retrieval (IR) community and to Web standards. In 2005, when I was still at AT&T, I moderated a panel in SIGMOD on "DBs and IR: Rethinking the Great Divide". We debated the question of rethinking data management system architectures to merge DB and IR technologies. I remember being lost in translation when trying to bridge the gap between Boolean queries and the need for ranked retrieval for

XML full-text queries. That confusion is so much clearer today.

Given how things turned out for XML, how do you feel about your own work and contributions?

My work on XML was a pivotal moment in my career. I started collaborating with various researchers and practitioners. In 2003, with Pat Case at the US Library of Congress, we wrote a W3C recommendation document where we designed use cases for XML full-text search, based on how Pat and her colleagues in the library accessed documents on the Web. Pat taught me that to work with people from other disciplines, we needed to learn to speak the same language, so to speak, and align our goals. With Jayavel Shanmugasundaram, we added full-text search primitives to XQuery and XPath, and our language, published in VLDB 2005, was integrated into a 2011 W3C recommendation.

While at Yahoo!, I had a chance to work with great IR experts, Ricardo Baeza Yates and Mounia Lalmas. XML allowed us to do a lot of work together. We published papers and gave tutorials at SIGIR and VLDB. It's interesting to see how the topics of our collaborations have evolved over time. It went from XML languages and the INEX (The INitiative for the Evaluation of XML retrieval)[1], to questions that were more fundamental, about how to integrate information retrieval and database techniques to solve XML retrieval questions[2], and that led to questions around accessing data on the Web[3].

My work on DB/IR integration culminated with a panel at VLDB 2007 with Alon Halevy and a SIGMOD Record paper where each panelist defended their statement: Alon defended the idea that the Web 2.0 is about helping the masses manage heterogeneous datasets collaboratively. Gerhard Weikum promoted the fact that the Web 2.0 is about content-production democracy and a data-quality crisis. Volker Markl and Donald Kossmann focused on how one could use database expertise to define mashups declaratively, and AnHai Doan outlined pressing database questions in the Web 2.0. On my part, I talked a lot about how to leverage social ties to find the right content to serve to the right user. And that had a long-lasting impact on the way I designed recommendation algorithms that made use of social behavior and social ties. Two years later, in 2010, I found myself working on data management questions for human-centric Web applications.

Crowdsourcing became a central topic in my work after I sat on a SIGMOD 2010 panel moderated by Michael Franklin on Crowds, clouds, and algorithms: exploring the human side of big data applications. Ten years later, in 2019, I co-organized a Shonan workshop titled "Imagine all the People and AI in the Future of Work." So, in hindsight, working on XML got me closer to people.

In all of this success, I assume there were ups and downs. We all have written papers that we feel were not appreciated enough. Is there any work you would like to talk about that didn't receive the attention it deserved?

In a way, my biggest failure is my greatest success. My most cited paper dates back to 2009 and is titled "Group recommendation: Semantics and efficiency." That work was about defining semantics for group recommendations and how to reconcile different users' perspectives, and how to do that efficiently in a dynamic fashion. That was work that I did with Senjuti Basu Roy, which in fact started a long collaboration with whom I still work.

That work really showed how data management solutions, materialization, indices, etc, can be used to design faster recommendation algorithms for individuals and for groups of people. I had great plans for that work: to serve as a basis for rethinking database architecture, models, and algorithms to handle groups, teams, and communities as first-class citizens. I thought these databases could serve as a backbone for building Web applications. When I joined CNRS, I recruited several colleagues in Grenoble to build SOCLE, a framework for data preparation in social applications, where we used several of those ideas that we had initially. I gave multiple tutorials at WWW, SIGMOD, and VLDB, wrote surveys, and collaborated with experts to add a visualization layer. We also had several accepted demonstrations in the viz and database communities. Despite that, I still feel highly unsatisfied because I did not bring it together into a single system. So many applications and user needs to reconcile, so many content retrieval and recommendation algorithms to bring together, and no one system to rule them all. I went through my paper titles, and my longest recurring words are "group/community" from 2007 to 2023! So

[1] Sihem Amer-Yahia and Mounia Lalmas: XML Search: Languages, INEX and Scoring. ACM SIGMOD Record, v 35(4): 16-23, 2006.

[2] Sihem Amer-Yahia, Ricardo Baeza-Yates, Mariano Consens, Mounia Lalmas: XML Retrieval: Integrated IR-DB Challenges and Solutions. SIGIR Tutorial, 2007.

[3] Sihem Amer-Yahia, Ricardo Baeza-Yates, Mariano P. Consens, Mounia Lalmas: XML Retrieval: DB/IR in theory, Web in practice. Proceedings of the VLDB, 1437-1438, 2007.

maybe I should not despair, and this may happen someday.

I believed we as a community would start rehauling DB systems to handle individuals and groups as first-class citizens, but that did not happen. We are proud, as a community, of building generic databases, and I believe we are a bit resentful (including myself) of building special-purpose databases. We need to talk about that. We need to talk more about our failures. Maybe another Failed Aspirations in Database Systems workshop would be great. I enjoyed FADS@VLDB 2017 very much, and I think I was not the only one who enjoyed it.

You have done a lot of work, besides your technical work, in terms of contributions to the community. You have initiated the Diversity, Equity, and Inclusion (DEI) initiative in the database community and chaired the DBDNI group for three years. Your DEI work has had a great impact. Were there particular events that motivated you to go down this path?

I just did not want to attend another women's lunch! In fact, Juliana Freire, the SIGMOD Executive chair at the time, asked me if I could have a DEI working group for SIGMOD. I told her I'll think about it… and then Jeffrey Ullmann received the 2020 Turing Award with Alfred V. Aho. A big controversy broke out on whether or not it was a good idea to celebrate Ullmann and his work as a community. At that time, I was the DEI chair for VLDB 2020, and I tried to put together a panel to discuss that, but I failed. Most people I contacted pushed back and expressed concern about being stigmatized when talking about that.

I felt we needed to take a step back and understand how to approach that kind of question and be less emotional about it. So, I went back to the social sciences and I discovered Gisèle Sapiro, a CNRS sociologist and historian who wrote a book titled "Can we separate the work from the author?". I felt relieved to find a scientist who could give us perspective. Gisèle asked me for all the material I could give her, and she researched Ullman's case and drew parallels with other scientists' cases. For her, that was like devising an algorithm for us. She told us that while ethics is a growing concern in scientific communities like ours, we are not the first ones to ask ourselves the question of the relationship between an author's ethics and their work. Since the feminist and civil rights movements, increasing attention has been paid to sexual harassment and discrimination in academia. Some scientists argue that authors who engage in unethical behavior should be cancelled or at least not rewarded for their work. In contrast, others contended that the work should be dissociated from its author. She outlined a plan on how to think about those questions. It helped to see that one could approach delicate questions constructively, instead of becoming emotional about them. Of course, that is only one aspect of DEI, and we understood that it is both important and fascinating. Today, the DEI initiative is about so much more, and I am glad it has evolved.

We are proud, as a community, of building generic databases, and I believe we are a bit resentful (including myself) of building special-purpose databases. We need to talk about that. We need to talk more about our failures.

Speaking of the initiative's evolution, you had unusual success in terms of co-sponsorship from multiple conferences and societies, which rarely happens with anything. Even recently, we had DBCares merging with the DEI initiative. Can you comment on this? How did you manage that?

I think all the stars were aligning – we had many great people interested in those matters. But let me first say that it is much easier to get things done at the level of individual research communities and then elevate them than at the level of organizations such as the ACM or IEEE. And for that, we are lucky to have the SIGMOD Executive Committee, the VLDB Endowment, the EDBT/ICDT Executive Committee, and TCDE, all of which adhered to the DEI initiative at different moments in time.

I looked into all the efforts that were happening in our conferences and felt there was potential to reduce redundancy, and sort of build a history together that goes beyond conference boundaries. I initially reached out to people in the DB community who have been involved in DEI events in the past with a proposal to run a meeting and pick their brains on the topic. I just did not want to repeat things, so we had a first meeting. Things grew rapidly from there. I also found that the SIGHCI community was at the forefront of DEI questions in 2020, and it served as a great inspiration for structuring our DEI initiative. I started running one-hour meetings every other month. Because it was during the pandemic, people felt excited about discussing topics that gave them a sense of purpose. As discussions unfolded, it became clear that we needed to define specific actions and designate ambassadors who could

advise individual conferences and help build continuity in our efforts.

Merging DBCares with DEI happened naturally when we started talking about the ethics action. Similarly, promoting the use of CLOSET for CoI detection became part of DEI because it is an ethics concern. In fact, SIGMOD and VLDB will cover the costs of hiring a software engineer for one year to develop a tool to help with PC formation and paper assignments based on CLOSET. I recently learned that the SIGSAC Executive Committee established the Committee on Preserving Professional Conduct and Academic Ethics (SIGSAC PROTECT) with the mission of providing a coordinated and timely response to emerging ethical concerns. Today, many other communities, including NeuRIPS and KDD, are reaching out to us to share our experience in setting up the DEI initiative.

Among the actions of the DEI initiative, there are many components (e.g., Reach out, Include, Organize, Support, Scout, and Coordinate), some of which you discussed. Is this vocabulary something that people are more broadly aware of? What are the most challenging tasks that you and your colleagues have addressed with a long-term impact?

This vocabulary is something I came up with because, for the first year, every time I thought about it before our next meeting, I would be doing things. For instance, *Scout* came up because I was really scouting for DEI events, trying to understand what other communities were doing, what was happening on universities' websites, etc. They were evolving and talking more about DEI. I said, "OK, maybe I should just coordinate the initiative and let my colleagues who were part of the initiative to scout, support, organize, include, and reach out". So that is how this vocabulary came up.

> ***... involving men in the DEI initiative and, more generally, in DEI events, was and still is the biggest challenge.***

We faced multiple challenges. In the reach out action, the idea was to design a single questionnaire and deploy it to every one of our conferences to understand the profiles of people attending our conferences and how attendance evolved over time. Of course, deploying those questionnaires came with the challenge of ensuring privacy – where would we store the data we gather about people? And before that, there were questions about how to design the questions. How do you ask about gender and sexual orientation? Would people be willing to provide that data?

We also encountered issues related to funding DEI efforts. One thing we realized is that we need to plan upfront and make sure conference organization proposals include a line for DEI events in their budgets, so that they are treated as a first-class concern in our events.

Thinking about what DEI means for journals and workshops is still ongoing. These and other challenges are discussed in our yearly SIGMOD Record reports. And while promoting DEI is honorable, enforcing DEI is not always easy – it is also risky, and we are not trained for that. Depending on what we are talking about, a harassment case or a CoI violation case, approaching it constructively without building stigma around the individuals involved is a hard question.

I risk falling into a cliché, but involving men in the DEI initiative and, more generally, in DEI events, was and still is the biggest challenge. DEI is a nurturing and caring activity; men and women have different ways of caring. We need to include both ways of caring in our DEI actions, and to do so, the initiative must include more men. In my career, several male colleagues have had a supportive role, including yourself, Jag. Thank you for that. We need our male colleagues to engage more in generous and empathetic behavior.

Beyond your own DEI efforts, what advice would you give women entering the field of DB research today?

First, I want more women. Please go into it! If you would like to do database work, please do! It is really fun, and the people are very nice. I love my colleagues!

I would like to give two pieces of advice to both women and men. The most important advice is to be aware of the fact that mentalities have changed. If you see something or experience discomfort, unease, or shock, you can talk to colleagues involved in the DEI initiative. You should not feel that it is your fault. If you still think it is your fault, it means that the initiative still has a long way to go. So, the initiative is there for you. The other advice is to realize that change does not happen by itself and that everyone is welcome to get involved in DEI efforts.

And to both men and women, an important thing is to have an activity outside of work to let off steam and explore other sides of your potential. Doing research is very personal and is highly rewarding when we succeed, and hard on us when we do not. Defining one's achievements solely through research is not a good idea. One needs to devise ways to compensate for failures. I do it by dancing and keeping in touch with friends around the world.

That is a good point to move on to more personal matters. You have worked in industry, academia, and research institutes. You have lived and worked in different countries. Do you have thoughts to share on how all these compare?

For a long time, research in academia and industry was conducted in very different ways. One thing that is quite unique in the industry, particularly in the Web industry, is that people with very different career paths and research areas are striving to achieve the same goal. That makes working together with other disciplines more natural. In academia, historically, boundaries between disciplines have been quite rigid. When I arrived at CNRS, it was the first time I started working in academia. Before that, I had been trained to work with people in other disciplines, and it took me a while to do that again since I joined CNRS.

Luckily, the advent of data science has been changing that. All the AI institutes are gathering people from different disciplines. Most researchers in research institutes such as CNRS and INRIA do not teach, as it is not required. That leaves plenty of time for them to chase funding for their research. However, they have less access to students. In France, we have mixed research units that co-locate University professors with CNRS and INRIA researchers. The lab I work in, Laboratoire d'Informatique de Grenoble, is a mixed research unit where we benefit from each other's perks. I spend more time raising funds for our research, and my colleagues spend more time convincing students to join us.

As for working in different countries, while chasing a paper deadline and other mundane activities we do as researchers feel the same everywhere, I must admit that the experience and sense of purpose change a great deal between places. In my opinion, the industry is rougher. When I was at Yahoo!, I worked very hard and was very excited about the research I was doing because I had access to great data and some of the brightest colleagues, but I never felt I belonged. I had tough and misogynistic bosses. I am not sure they or I were fully aware of that. Luckily, I was very excited about my research. Also, attending conferences and seeing my friends and colleagues was a great consolation, and living in NYC allowed me to make great friends and practice my dancing. That's a big part of my life. I hope things are different today in the industry. I can't really tell.

When I joined QCRI, in Qatar, Ahmed Elmagarmid suggested we put together a mentorship program for undergraduate students to come and spend time in the lab, participate in research projects, and get ready to apply for grad school outside of Qatar, since there were no graduate programs there. I was very surprised to see so many women sign up for that program. It turns out there are many more women than men who study Math and Computer Science in the Middle East and North Africa region. I come from North Africa, and I did not even know that. Most of the interns we had ended up doing a PhD in prestigious universities in the UK and the US. That really gives you a great sense of purpose.

Talking about the sense of purpose, what have you found to be the most rewarding in your work life?

The most rewarding thing is learning from other research communities, both in Computer Science and in other fields. Lately, I started collaborating with Education scientists, and I am discovering different theories on how people learn alone and with others. Some of those theories are making it into my recent research. It's amazing to have such freedom. And, probably the most important thing is to meet people from all over the world who think differently, are smart, hardworking, and ambitious, and among them, beautiful people like Divesh Srivastava and Tova Milo, who lift you up.

You have enjoyed dancing all your life. Can you tell us something about that?

I grew up in Algeria, and I was around 4 years old when I took my first dance class. Later, I became a member of the Algerian National Ballet. There was a point in my life, when I was in high school, where I was given a choice of dancing more and doing less math, or continuing to do what I was doing. It was a tough decision. I ended up dancing less, and I found locations to dance less professionally.

To me, classical ballet is like Boolean queries in relational databases. Let me attempt to do this parallel. Classical ballet is very well defined. There is this one movement, you have to do it the way it is dictated: your legs are either plié or tendu. The former is a basic bending of the knees while keeping the heels on the ground. In the latter, the legs are fully stretched. So your legs are either bent or stretched. When I arrived in France as a student, I started Modern Jazz. In Jazz, a tendu is a fluid movement that travels through checkpoints without stopping. Your legs are never totally tendu or totally plié. They're always in between. It's more like IR. Everything is a potential answer to a search with a score. When I moved to NYC, I learned to dance Simonson's Jazz, an organic approach to movement that prepares the body to dance in a way that complies with your anatomy. In Grenoble, I've been dancing Horton Jazz, which focuses on stretching in opposite directions and smoothly connecting flat backs and lateral stretches, tilt lines, and lunges.

I feel like my work life has gone through that kind of evolution. In fact, my whole life has gone through that evolution. Living and working in different places taught me to better understand who I am and what I seek, build smooth transitions, recognize what I like in a place and be grateful for it, and approach my life and choices holistically.

While specialization in science has contributed to remarkable progress, the separation between fields that aim to maintain their distinctiveness constitutes an obstacle to innovation and collaborative efforts.

That is an amazing parallel! To close up our conversation, let's go back to technical stuff. How do you see the future of data management research, and what is the next pressing challenge for us as a community?

We are the data experts, and we know how to deal with data. We have growing amounts of data, including data about people, and that should really help us to understand how to care more about people. To do that, we need to take a step back, understand what "Transdisciplinarity" means, and focus on integrating intellectual frameworks that transcend individual disciplinary viewpoints. Conceptual frameworks from different fields can provide a broader perspective in both research and practice. For instance, in Positive Psychology, there are several theories that can be applied to the field of AI & Well Being, and that have so much to teach us in terms of paying more attention to people when building human-facing and human-caring DB systems. The work on fairness is going in that direction, and I think we can do more by attempting to answer other fundamental questions, such as "How do we capture experience, satisfaction, and frustration that users experience when interacting with data? How to devise data processing algorithms that optimize for positive feelings? The good news is that there are many theories, such as the Flow Theory in Psychology and the Self-determination Theory, that can help us.

From an intellectual standpoint, advancing research in one's field can be significantly influenced by other disciplines' theories, concepts, and methodologies. While specialization in science has contributed to remarkable progress, the separation between fields that aim to maintain their distinctiveness constitutes an obstacle to innovation and collaborative efforts. Practically speaking, the challenges currently confronting our world do not align neatly with academic disciplines; instead, they are increasingly complex, chaotic, and interrelated, and humans are in the middle of that. Consequently, there is a growing recognition of the necessity for a more comprehensive and integrated approach to understanding these multifaceted issues. That can only be achieved by transcending disciplinary boundaries. This situation further supports the argument for reforming educational practices and advocating for a more cohesive and integrated curriculum in our universities.

Thank you, Sihem, that is a wonderful place to end.

Thank you, Jag, once more!

Diversity, Equity and Inclusion Activities in Database Conferences: A 2024 Report

https://dbdni.github.io

Nelly Barret, Sourav S Bhowmick, Angela Bonifati, Barbara Catania, Stratos Idreos, Ekaterini Ioannou, Madhulika Mohanty, Sana Sellami, Roee Shraga, Utku Sirin, Juno Steegmans, Pınar Tözün, Soror Sahri, Genoveva Vargas-Solar

1. THE DEI@DB INITIATIVE

The database community's Diversity, Equity, and Inclusion (DEI) initiative began in 2020 as the Diversity/Inclusion initiative [1]. This report highlights our activities from 2024. Our goal as a community is to make all DB conference attendees feel included, regardless of their scientific views or personal backgrounds. As a leadership team, the DEI group supports DEI chairs across conferences, preserves institutional memory of DEI efforts, shapes a shared vision, and fosters collaboration to advance inclusion. These efforts are carried out by core members (Figure 1) and liaisons from each conference's executive committee (Figure 2). The initiative was relaunched in January 2024 with a new structure based on five key actions: **COORDINATE**, to support collaboration between core members, liaisons, and DEI chairs; **SCOUT**, to gather best DEI practices from other communities; **ETHICS**, to create and promote ethical guidelines for writing and reviewing; **MEDIA**, to collect and share digital content from DEI@DB events [4]; and **DIVERSIFY**, to analyze data on diversity, accessibility, and the adoption of DEI principles in research and academia. DBCARES[1] is now officially part of the DEI initiative. The mission of DBCARES is to create an inclusive and diverse Database community with zero tolerance for abuse, discrimination, or harassment. As part of this integration, we unified the Code of Ethics and introduced clear guidelines for DB conference organizers. Several conferences—including SIGMOD, VLDB, ICDE, and EDBT—continued using CLOSET [2] to ensure fair and transparent reviewer assignments.

The **SUPPORT**, **INCLUDE**, **INFORM**, **ORGANIZE**, and **REACH OUT** actions have now become standard practice, and since 2024, they are no longer managed by the DEI initiative. To ensure early financial planning for DEI conference activities, **SUPPORT** is now the responsibility of conference organizers. At the same time, **INCLUDE**, **ORGANIZE**, and **REACH OUT** fall under the direct responsibility of DEI chairs at each conference.

What did we achieve this year? The database community's engagement with DEI initiatives in 2024 has been highly encouraging. Shared experiences across DB conferences have enabled the scaling and enrichment of DEI activities throughout the year. As part of this progress,

[1]https://dbdni.github.io/pages/dbcares.html

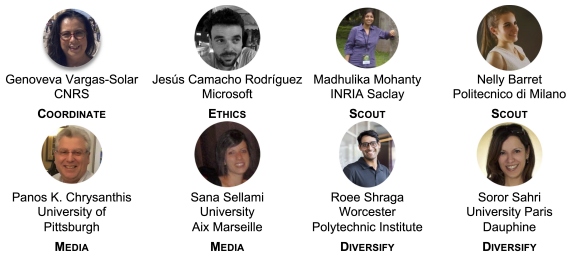

Figure 1: DEI@DB Core Members

Figure 2: DEI@DB Liaisons Members

the **SCOUT** initiative introduced DEI checklists for authors and reviewers to be embedded in submission and review forms, helping assess alignment with DEI principles. The Author Checklist promotes inclusive language and diverse visuals while avoiding stereotypes and oppressive terms. The Reviewer Checklist fosters respectful, detailed, and constructive feedback. Based on VLDB's DEI practices, these checklists will be gradually adopted in line with each conference's timeline. **ADBIS 2025** will pilot the initiative, setting an example for others. We will monitor, refine, and share outcomes to foster a more inclusive research culture. Several conferences launched caregiving and wellness initiatives, including childcare spaces with dedicated programming (e.g., SIGMOD and SIGSPATIAL), and quiet or wellness areas for attendees in need of rest or decompression (e.g., ADBIS). These services were well received, with positive attendee feedback. However, organizers emphasized the need for clearer communication to ensure such resources are known and factored into travel planning. To expand the impact of our efforts, we have also reviewed inclusion programs from other ACM conferences, such as **ACM FAccT**, to explore a broader range of support mechanisms. In 2025, we will assess and adapt

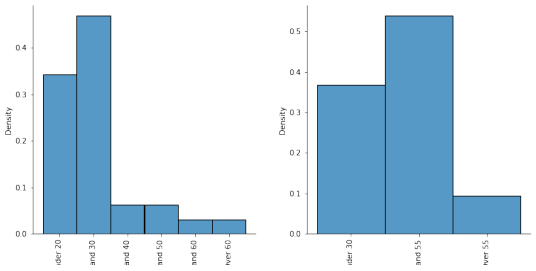

Figure 3: Age distribution of survey responders.

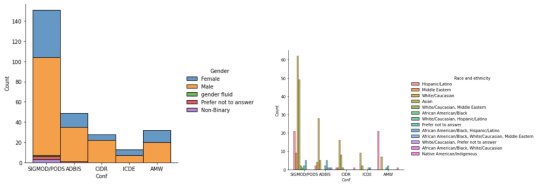

Figure 4: Race, ethnicity and gender distribution.

promising models to suit the needs and dynamics of the DB community.

2024 DEI statistics. A key goal of DEI@DB is to better understand our community, identify areas for improvement, and evaluate the impact of our initiatives. To support this, we conducted surveys at CIDR, ICDE, ADBIS, SIGMOD/PODS, and AMW, with 30, 13, 48, 155, and 32 respondents, respectively. Results are aggregated across these conferences. Participants were primarily from academia (26%), industry (12%), or both (33%), with students making up 28%. In contrast to 2022 and 2023 (which included hybrid events), all surveyed conferences in 2024 were in-person. Figure 3 shows participant age distribution: initially, most respondents were between 30–55 years old. Later surveys used a more fine-grained age range and showed a shift toward the 20–30 group—mainly due to the updated question being implemented only at AMW. This refined format is now standard for 2025 surveys, and we plan to revisit the distribution as more data becomes available.

The plot on the right of Figure 4 focuses on race and ethnicity: a clear dominance of participants identifying as White / Caucasian is observed across all conferences, particularly at SIGMOD/PODS and ADBIS. Representation from groups such as Asian, Hispanic / Latino, Middle Eastern, and African American / Black appears significantly lower. However, some multiracial and intersectional identities (e.g., White / Caucasian & Hispanic / Latino) are also captured. Notably, a small but consistent group of respondents selected "Prefer not to answer," suggesting privacy concerns or limitations in identity categories. The plot on the left, examines gender distribution. Roughly 32% of respondents (about 85 individuals) identified as female, and 9.5% as LGBTQ+, consistent with previous years. The plot shows a marked gender imbalance: male participants make up the majority in every conference, with female participation forming the second largest group. Non-

Figure 5: DEI Chairs of 2024 DB Conferences

binary and gender-fluid individuals are present in minimal numbers, as are those who prefer not to disclose their gender. SIGMOD/PODS again shows the highest overall participation, but the gender imbalance is consistent across all venues. Together, these plots highlight the ongoing diversity gaps in the database research community, reinforcing the importance of inclusive outreach, equitable access, and active efforts to support underrepresented groups in academic events and conference participation.

Survey results show that the most requested topics for DEI sessions are research-related issues (identifying topics, setting goals, and defining success), followed by work-life balance and mentorship. At SIGMOD, 14% of respondents reported that their research incorporates DEI concerns, which is a promising evolution sign. Suggestions for improvement included maintaining double-anonymous review, increasing diversity in program committees, supporting students and parents with infants, lowering registration fees, expanding mentorship opportunities, and easing visa processes for conference travel. A post-conference survey with 67 responses revealed that 95% of attendees had no visual or auditory difficulties, 86% were satisfied with the food, and 97% found the venue easy to navigate. Notably, 96% of respondents described the SIGMOD/PODS community as supportive.

2. DEI@DB CONFERENCES 2024

Figure 5 reports DEI chairs of individual conferences in 2024. In chronological order, we briefly report on the various activities at past and future database conferences.

DEI@CIDR. Utku Sirin, Stratos Idreos, & Pinar Tözün served as DEI co-chairs. The DEI program featured a mentoring initiative that paired junior and senior attendees for 10 one-on-one sessions, allowing each pair to choose their preferred meeting time and format. Additionally, James Hamilton hosted a group mentoring session with students. To organize these activities, the co-chairs reached out to CIDR participants shortly before the conference to identify mentors and mentees.

DEI@EDBT/ICDT. Julia Stoyanovich & Benny Kimelfeld were the DEI co-chairs. They organized an interactive session titled "Unfinished Comics for Inclusive Commu-

nication about Data Management in Research and Practice", encouraging participants to reflect on diversity and inclusion through personal experiences and explore creative ways to make data management more welcoming and accessible. They also presented an "Interactive Tutorial on Giving Inaccessible, Unclear, and Boring Presentations", designed to raise awareness of presentation quality with a focus on accessibility and diverse audiences.

DEI@ICDE. Ekaterini Ioannou and Pinar Tözün (ITU) served as DEI co-chairs. They organized a session where Prof. Alexander Serebrenik gave a talk on "Diversity, Inclusion, and Software", a lunch hour that brought together some of the senior members of the community with the junior members focusing on mentoring, and financial support to two students and one junior faculty member for their conference attendance.

DEI@MDM. Bettina Berendt and Vana Kalogeraki were the DEI chairs. The program featured a keynote by Pinar Tözün titled "Data Processing at the Edge: From Satellites to Earth". A DEI grant program supported the participation of students and early-career researchers from underrepresented communities at MDM 2024. The grants were possible through funding from IEEE TCDE, the Emeralds Horizon EU project, and the SoBigData++ Horizon 2020 project.

DEI@SIGMOD. Aidan Hogan and Jesús Camacho Rodríguez served as DEI co-chairs. SIGMOD/PODS activities included a "Birds of a Feather" session, to share DEI statistics, feedback, and improvement strategies for the conference and related events. The DEI panel, "Global Voices in Data: Navigating Responsible Management and Processing with Diverse Perspectives"[2], brought together voices from the Global North and South to discuss key factors for fostering responsible and accountable database research. Additional actions included sharing an Anti-Harassment Policy and running a post conference survey to collect attendee feedback on their experience and views on DEI at the event.

DEI@VLDB. Avrilia Floratou served as the DEI chair, promoting awareness and inclusivity throughout the conference. Key initiatives included publishing guidelines on writing and presenting research contributions with DEI considerations and actively communicating the code of conduct to ensure an inclusive and respectful environment.

DEI@ADBIS. Barbara Catania and Genoveva Vargas-Solar served as DEI co-chairs. The DEI program included a hybrid panel titled *"New masculinities: Do we need muscles in the lab?"*[3], and a keynote by Rita Bencivenga of University of Genoa on *"Gender+ and Intersectionality in EU projects."* A coordinated effort between the DEI and Doctoral Consortium led to joint activities. These included a hands-on data science session, *"DEI Perspectives in Data-Driven Experiments,"* led by Barbara Catania and Martina Brocchi a PhD student, and an in-person mentoring session to foster interaction among PhD students and early-career researchers. A dedicated privacy room in the venue's library provided a quiet and secure space for personal use. The EasyChair review form included a DEI criterion, encouraging reviewers to consider submissions through a DEI lens. To support inclusive participation, DEI co-chairs highlighted location-specific considerations and communicated the code of conduct, DEI guidelines, and announcements to organizers, keynote speakers, and authors—promoting respectful and inclusive engagement. A kakemono [4] summarizing DEI goals and achievements was displayed at the registration desk and included in the welcome materials to increase visibility.

DEI@SIGSPATIAL. Yao-Yi Chiang, Jina Kim, and Verity Hardgrove served as DEI co-chairs for SIGSPATIAL, leading initiatives to promote awareness and inclusivity throughout the conference. Key efforts included publishing guidelines on writing and presenting research with DEI considerations and actively communicating the code of conduct to foster an inclusive and respectful environment. Additionally, the conference provided caregiving facilities for attendees with children and proposed a travel awards program to support participants with travel and conference expenses. The U.S. National Science Foundation (NSF), conference sponsors, and ACM SIGSPATIAL funded the grants.

DEI@AMW. Genoveva Vargas-Solar and Javier-Alfonso Espinosa-Oviedo served as DEI co-chairs, shaping the scientific program through an intersectional lens considering gender, career stage, geography, nationality, and institutional background. The actions included offering health-conscious, sugar-free meals made with organic and locally sourced ingredients, and cultural breaks such as visits to biodiversity sites, contemporary art museums, and dance sessions to balance long sedentary activities. To support student participation, a grant program was launched, with 90% of recipients being graduate students from Indigenous and low-income backgrounds studying at public universities in Mexico. Online access was also provided to facilitate participation from students across South America. The grant program was funded by the VLDB Endowment, generous speakers, and contributions from participating institutions.

3. COI MANAGEMENT

As of 2024, the automated detection and management of conflicts of interest (COIs) has become a standard practice across major database conferences, including SIGMOD, VLDB, ICDE, and EDBT. These conferences have collectively adopted CLOSET [2] as a core tool for managing COIs, recognizing its effectiveness in automating this critical aspect of the review process. Notably, VLDB 2024 piloted an enhanced workflow by integrating results of CLOSET into the CMT submission system. In this pilot, the system pre-populated potential COIs for each author based on coauthorship data

[2] http://vargas-solar.com/dei-sigmod-pods-panel/

[3] http://vargas-solar.com/adbis-dei/dei-panel/

[4] A kakemono is a vertical hanging scroll—traditionally Japanese art for exhibitions.

and prompted authors to verify and confirm the accuracy of the detected conflicts, thereby streamlining the process and reducing manual input errors. However, it appears that a significant number of authors did not actually review or confirm their listed COIs.

We also observe a noticeable gap between the community's stance on penalties for under-declared COIs, such as desk rejections, and how these policies are actually enforced across major data management conferences. A recent community-wide survey [5] revealed strong support for strict penalties in such cases. However, enforcement remains inconsistent and unclear: while some venues impose certain penalties, others do not apply any at all.

4. GOING FORWARD

Job descriptions. We are working with the ACM to ensure the job descriptions of DEI members and chairs are aligned with the ACM policies. ACM sets global principles and enforcement pathways; our DBDEI initiative delivers domain-specific execution (surveys, templates, DB-conference workflows, and software).

COIs. While major conferences have adopted CLOSET to manage conflicts of interest (COIs), several others, such as CIDR and SIGSPATIAL, still rely on traditional tools that lack the same transparency and precision. More importantly, enforcement of penalties for under-reported or misrepresented COIs remains inconsistent and does not meet the expectations of the research community. This gap calls for stronger action: executive committees should lead efforts to align COI enforcement with community standards. This means adopting reliable tools like CLOSET and defining clear, consistent policies and consequences to ensure COI rules are applied fairly across all conferences.

Checklist for DEI Writing. We introduced a DEI writing checklist at ADBIS 2025 to encourage inclusive and responsible research communication. Our goal is to adopt this checklist across other database conferences and continuously refine it based on community feedback. We also plan to conduct surveys to assess its effectiveness and impact over time.

MEDIA Action. We launched the channel `DEI-DB-MEDIA` on YouTube [5] to centralize and share recordings of DEI related events, talks, panels, briefs, and workshops, organized by conference and year. To support this, we distributed a Google Form in 2024 to collect slides, videos, and links from that year's DEI activities. We also invited 2025 DEI co-chairs to contribute materials after their conferences. The goal is to build a shared archive of DEI programs across the database community, helping promote best practices, inspire new ideas, and provide educational resources. We also plan to enhance the site with summaries of each initiative, highlighting outcomes, common strategies, and lessons learned.

ETHICS action. We are working on establishing and promoting ethics guidelines for publications, similar to other efforts [3]. This involves creating a living document specifying major ethical aspects that authors and reviewers should consider. To enhance inclusion, we plan to compile a set of guidelines for session chairs, presenters, and participants for handling panels and Q&A. This action will unify the guidelines used by the author, reviewer, and presenter at individual conferences. It will also generalise the "checkbox" to flag institutional representation and SCOUTING action.

DIVERSIFY Action. We developed a diversity survey to assess representation and inclusion across database (DB) conferences. The survey covers leadership diversity, career stage and institutional representation, and accessibility features such as ramps, childcare, and accessible materials. It also evaluates DEI content in the program, available funding, and follow-up efforts after the conference. Responses from DEI chairs showed that gender diversity, varied career stages, and institutional representation were generally well addressed. Ethnic and cultural diversity in leadership also showed improvement. Accessibility measures were partially implemented, and support services like childcare were limited. DEI sessions were held at SIGMOD, ICDE, and ADBIS, with post-conference initiatives already active at SIGMOD.

Looking ahead, we plan to explore new actions based on community interest, including **education**, **sustainability**, **mind the gap** to address gender disparities in database research, and **amplify**, a mentoring initiative to help research groups strengthen their work and aim for top-tier publications and funding.

[5] `https://tinyurl.com/2v4ed98n`

5. REFERENCES

[1] Sihem Amer-Yahia, Yael Amsterdamer, Sourav S. Bhowmick, Angela Bonifati, Philippe Bonnet, Renata Borovica-Gajic, Barbara Catania, Tania Cerquitelli, Silvia Chiusano, Panos K. Chrysanthis, Carlo Curino, Jérôme Darmont, Amr El Abbadi, Avrilia Floratou, Juliana Freire, Alekh Jindal, Vana Kalogeraki, Georgia Koutrika, Arun Kumar, Sujaya Maiyya, Alexandra Meliou, Madhulika Mohanty, Felix Naumann, Nele Sina Noack, Fatma Özcan, Liat Peterfreund, Wenny Rahayu, Wang-Chiew Tan, Yuanyuan Tian, Pinar Tözün, Genoveva Vargas-Solar, Neeraja J. Yadwadkar, and Meihui Zhang. Diversity and inclusion activities in database conferences: A 2021 report. *SIGMOD Rec.*, 51(2):69–73, 2022.

[2] Sourav S. Bhowmick. CLOSET: data-driven COI detection and management in peer-review venues. *Commun. ACM*, 66(7):70–71, 2023.

[3] NeurIPS Foundation. Ethics guidelines. `https://nips.cc/public/EthicsGuidelines`.

[4] DB DEI Initiative. DB DEI Materials. `https://dbdni.github.io/#materials`.

[5] Alexandra Meliou, Sourav S. Bhowmick, Karl Aberer, Divy Agrawal, Angela Bonifati, Vanessa Braganholo, Floris Geerts, Wolfgang Lehner, and Divesh Srivastava. Peer-reviewing processes and incentives: Data management community survey results. *SIGMOD Rec.*, 52(4):41–46, 2023.