A mapping study on query optimization using Machine Learning techniques

Matina Korkontzelou

September 2023

1 Introduction

1.1 The importance of query optimization

Query optimization is crucial for efficient relational database management systems. It examines multiple strategies to determine the best way to execute a given query. Although there can be many approaches to run a single query, the aim is to pick the one that uses system resources most efficiently, reduces costs, and delivers quick, accurate results. In an era where data is vast and time is of the essence, effective query optimization is vital. It not only ensures that databases run smoothly and cost-effectively but also gives organizations a competitive edge by offering faster, more reliable results.

1.2 The need of Machine Learning in query optimization

Machine Learning, an essential facet of Artificial Intelligence, harnesses data and prior knowledge to discern patterns, aiming to emulate human learning. Its overarching ambition is to derive insights and predictions with limited human input. Traditionally, query optimizers have selected query plans based on cost models grounded in cardinality estimations. However, this approach isn't infallible. Misjudgments in these estimations frequently result in less than optimal query plans. Moreover, the onus largely falls on database administrators to meticulously refine these cost models to ensure efficient query execution, a process that can be labor-intensive and prone to errors. This underscores a pressing need for query optimizers that adaptively learn from data. Incorporating Machine Learning can usher in a new generation of optimizers, tailor-made for sophisticated analytical databases, that are both more accurate and self-evolving.

1.3 The challenges of Machine Learning in query optimization

Employing machine learning for query optimization presents discernible challenges. Paramount among these is the acquisition of quality training data,

integral for accurate model predictions. Given the dynamic nature of databases, models necessitate periodic recalibrations, a process that can be resource-intensive. The phenomena of overfitting and underfitting further complicate the optimization process, requiring meticulous attention to the granularity of training data. Moreover, the inherent opacity of machine learning models, often perceived as 'black boxes', can pose challenges for database administrators in terms of interpretability. However, in juxtaposition with traditional methodologies, machine learning offers unparalleled adaptability, the capacity to process and learn from extensive data sets, and the ability to perpetually refine its methods. Consequently, even amidst these challenges, machine learning emerges as an advanced and preferable approach for query optimization.

1.4 The need for a structured map of the body of knowledge

A body of literature in the area of query optimization using Machine Learning does not exit. The goal of this paper is to provide a "map of the terrain" of the literature in query optimization using machine learning evolution that covers the most recent years, via a principled and organized protocol for compiling the body of literature and with answers to specific research questions around the trends, venues, and research problems that the scientific community addresses in the area of query optimization using machine learning. To this end, in this paper, we provide the plan, execution details, results, and threats to validity for a systematic mapping study in the area of query optimization using machine learning that fulfills this goal.

Roadmap

In Section 2, we provide the background on what systematic mapping studies are. In Section 3, we detail the protocol of the systematic mapping that we have performed. In Section 4, we present the execution of the designed protocol and the construction of a corpus of documents for further analysis. In Section 5, we present the results of the analysis that we performed on the basis of the aforementioned plan. In Section 6, we discuss threats to validity. Finally, in Section 7, we summarize our findings and conclude our discussion.

2 Background and Related work

2.1 Systematic Mapping studies in Computer Science

2.1.1 What is a Systematic mapping study?

A systematic mapping study is a structured overview of the publications in a certain scientific topic with the aim to give a broad overview of the quantity and type of the research being conducted in a scientific area, but without delving into the deep details of methods and concrete answers to research questions [5] , [8] , [2] .

The results of a systematic mapping study is, broadly speaking, a map of the "terrain" of a broad research topic (as in our case "query optimization using

machine learning") in order to identify the trends, venues, types of publications and broad research questions concerning the research community. The discussion is not going in great depths, or deep analyses of the methods used or answers given to the research questions by the research community. Providing an overview of the opinions, consensus, digressions, findings for a narrow research question is the job of a *systematic review*, a close cousin of a systematic mapping study, whose task is to inform researchers on the current state of the art and practice for a very specific research question. Thus, typically, systematic maps precede systematic reviews by charting the potential research questions that the scientific community is concerned with. *A common feature both "cousins" share, however, is that they are conducted on the basis of a very specific protocol* and with the goal to include as many as possible publications of the area they survey, in a principle, automatized, and strictly guided manner, in order to avoid personal biases, or accidental omissions of research works.

2.1.2 Why bother with Systematic Mapping Studies?

As time passes, research results ("primary studies") within a certain broad area are produced, and the number monotonically grows. Therefore, in order to allow researchers orient themselves with the problems and methods of this broad research area there is a need to classify the publications firstly, and, after that, to further structure the results of concrete research questions [8]. This allows, first, to gain an overview of the area, and second to delve into the general state of affairs for specific research questions, both for the newcomer, but also for the researchers who have been working in a certain area for long.

Systematic Mapping Studies also provide us with (a) a better understanding of trends in the particular research area, and, (b) the possibility of identifying gaps and missed opportunities by the existing primary studies.

2.1.3 Guidelines and protocols for Systematic Mapping Studies

Systematic mapping studies have a wide use in medical research but are not so frequent in software engineering, and -to the best of our knowledge- completely absent from the data engineering literature. One particular difficulty has to do with the lack of very specific guidelines about how to apply this type of research survey in a principled manner [8], [2]. The most celebrated set of guidelines is the one provided from Kitchenham and Charters in 2007 [5]; however, these guidelines are focused mainly on performing data extraction and analysis for Systematic Literature Reviews and not so much in protocols and steps for conducting a Systematic Mapping Study. To the best of our knowledge, the most comprehensive set of guidelines for performing Systematic Mapping Studies is a fairly recent paper by Petersen et al., in 2015 [2]. In what follows, we adopt the general method of [2] for the planning and execution of our mapping study. We report on the method followed in Figure 1.

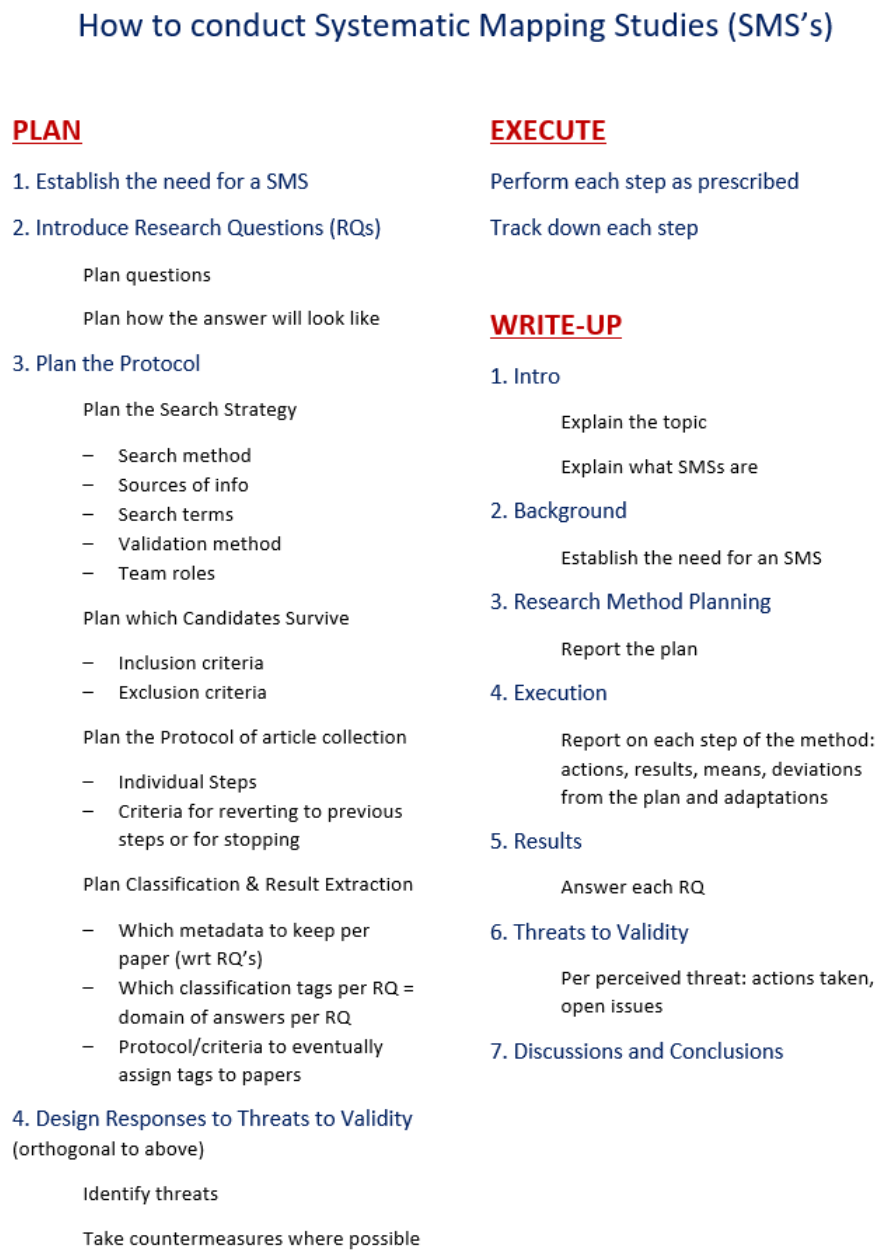

Figure 1
The general workflow of a Systematic Mapping Study

2.2 Existing secondary studies in Query Optimization using Machine Learning techniques

In this section, we present a review of recent secondary studies, specifically surveys and mapping studies, within the domain of query optimization leveraging Machine Learning techniques. Contrary to primary studies, which document original and novel technical findings, secondary studies focus on collating and analyzing existing literature. It is noteworthy that our rigorous exploration did not yield any studies analogous to our current undertaking.

3 Research Method

The main goal of this mapping study is to
• identify, and,
• structure
the characteristics of the research literature in the area of query optimization using machine learning, and particularly,
• the research questions and methods,
• the type of data,
• the publication venues,
• the publication types (journal articles, conference papers, summaries and surveys, tutorials, or other),
• the time trend (amount of publications per year)

All the above constitute the *metadata* that we will retain for each retrieved publication.

3.1 Research Questions

The specific research questions that we pose are delineated in Table 1.

Table 1
Research Questions

RQ	Research Question
1	What is the annual amount of publications and the overall trend?
2	What is the breakdown of the publications in terms of venues and types?
3	In which domain or application area does the paper investigate the utilization of machine learning techniques for query optimization?
4	What are the primary research questions and solution methods proposed in this paper concerning query optimization using machine learning?
5	Does the paper present a specific architecture for a machine-learning-based query optimizer, or does it theorize on optimal structures without detailing a concrete architecture?
6	What machine learning technique does the paper employ for query optimization, and what outcomes are achieved through its implementation?
7	How is the training of this machine learning model accomplished in the study?
8	Does the paper offer any source code or implementation details for the proposed query optimizer?
9	On which database systems has the optimization been tested or implemented?
10	Which methods for search space pruning are highlighted or employed in this study?
11	How does the query optimizer proposed in this paper compare with existing query optimizers in terms of databases used and overall performance?
12	What are the identified limitations or drawbacks of the machine learning technique proposed in this study for query optimization?
13	Based on the findings and discussions in the paper, what avenues for further research are suggested or appear promising?

The first two questions are *topic-independent* and can be applied to any surveyed research area. The two independent questions concern the "when?" and "where"? questions, and their breakdown can be used orthogonally to the following topic-dependent, research questions. The result of each question involves a table of frequencies.

The next questions are topic dependent.

RQ3 concerns the specific domain or application area within which the papers examine the utilization of machine learning techniques for query optimization.

RQ4 concerns the particular general problems and the types of methods employed by the papers to attack the problems (e.g., the papers can perform a case study, or can propose a method or algorithm, etc).

RQ5 seeks to identify if the papers present a specific architecture for a machine-learning-based query optimizer or if they merely theorize about potential structures without detailed explanations.

RQ6 focuses on identifying the machine learning techniques used in query optimization and assessing their effectiveness. It aims to understand the range of methods applied and the practical or theoretical results they yield. Answering this provides insights into the current best practices, innovative approaches, and research gaps in query optimization using machine learning.

RQ7 is primarily concerned with understanding the methodology used for training the machine learning model. This includes insights into the rigor of the training process, the nature of the training data, specific training techniques or algorithms used, evaluation metrics, and hyper parameter tuning methods. Overall, the aim is to assess the robustness and validity of the machine learning approach presented in each study.

RQ8 aims to ascertain whether the collected papers provide source code or implementation details for their respective query optimizers. This inquiry underscores the collective emphasis on reproducibility and practical applicability across multiple studies.

RQ9 seeks to identify the variety and scope of database systems addressed across the collection of papers. The concern is to understand the breadth of application and potentially gauge the universality or specificity of the optimization techniques. When examining multiple studies, this provides insight into commonalities or gaps in the research landscape. This question aims to map the range of database systems explored in the collective papers to discern prevalent trends or overlooked areas.

RQ10 aims to identify the techniques utilized for search space pruning across the array of papers. The central concern is discerning which methods are frequently employed, recognizing any emerging patterns, innovations, or gaps in the collective body of research. Summarized, the question seeks to understand prevalent and varied search space pruning techniques highlighted across multiple studies.

RQ11 assesses how proposed query optimizers in various papers compare to existing standards, considering database compatibility and performance. Essentially, it aims to pin-

point advancements or gaps in optimization techniques across the studied literature.

RQ12 probes the recognized limitations or flaws of machine learning methods for query optimization across a multitude of papers. The central concern is to collate common challenges or shortcomings cited in these techniques. In essence, it seeks to compile a comprehensive understanding of the constraints and issues in machine learning-based query optimization from the collective research.

RQ13 seeks to extract future research directions or promising areas highlighted across several papers. The core intent is to map out emerging trends or underexplored avenues in the domain of query optimization. Briefly, it aims to collate collective insights on potential next steps in the field from the array of studies.

3.2 Search Strategy

3.2.1 Search Method

The first thing to clearly outline when planning the systematic mapping study is define the appropriate search method. A search method is an efficient way to find the metadata and identification information about the papers that pertain to the research questions we want to answer. Clearly, the integrity and completeness of the collected information are the keys to support the validity of the study.There are different types of search methods that are analyzed below, where we follow the excellent classification of [1]

Types of Search methods

• Automated Search:
Automated search is a search method that uses resources like digital libraries and indexing systems [1].
• Manual Search :
Manual search also known as handsearching is a search method that involves a manual page-by-page examination of the entire contents of a journal issue or conference. It is very useful for finding studies that appear in journal supplements or special editions that might not make it into databases [1].
• Snowballing :
Snowballing (also known as "citation searching" or "pearl growing") is based on starting from a key set of reference papers and attractively follow their references. Backward snowballing means looking for references that are cited by the key reference papers, or by the papers added to the current corpus in previous iterations. Forwards snowballing means looking for references that cite key reference papers, or the papers added to the current corpus in previous iterations. This can be obtained by querying publicly available paper collections like Google Scholar on the papers citing a certain publication. [1].
It is quite possible, that such methods are combined in order to enlarge the completeness of the collected corpus. The success of such methods, nowadays, depends heavily on the choice of public repository that is used to assist the retrieval of information, the queries posed to it, as well as the checks and balances used to mitigate the risk of failing to include important literature. We detail such considerations in the following subsections.

3.2.2 Search Sources

In this subsection, we analyze the digital libraries that could be used for our search in order to find the material related with query optimization using ML, as well as our design decisions on how to conduct the search.
Nowadays, there are many online journal and research databases which some of them presented below:

Database sources and web links	
ACM digital library	https://dl.acm.org
IEEE digital library	https://ieeexplore.ieee.org
Arxiv	https://arxiv.org
Google Scholar	https://scholar.google.com
Scopus	https://www.scopus.com/home.uri
DBLP Computer Science Bibliography	https://dblp.org

The DBLP Computer Science Bibliography is a joint service of Schloss Dagstuhl - Leibniz Center for Informatics and the University of Trier. It is supported by a team of scientists from Schloss Dagstuhl and supported by an advisory board of internationally renowned experts. DBLP collects information via scripts and personal collaboration with editors and is characterized by the combination of (a) high quality and (b) breadth of its collection that covers "major computer science publications" – for example, as the site mentions "As of January 2019, dblp indexes over 4.4 million publications, published by more than 2.2 million authors. To this end, dblp indexes about than 40,000 journal volumes, more than 39,000 conference and workshop proceedings, and more than 80,000 monographs"[1].

Moreover dblp comes with a dedicated search facility `https://dblp.org/search/` that returns all the papers that it indexes, fulfilling a set of criteria for the title, venue and author names, as well as customization options (see `https://dblp.org/faq/13501473.html`).

The advantage of DBLP compared to Arxiv, Google Scholar, and the ACM and IEEE digital libraries is that it practically encompasses them all under a single umbrella, includes all the major venus of publication in Computer Science, avoids -to a large extent- non-peer reviewed publications and is supported by data cleaning and data integrity efforts by its staff to ensure data quality. Therefore, we believe that DBLP on its own, is sufficient to support our mapping study for providing the metadata for papers related with query optimization using machine learning techniques. Once the metadata are collected, the retrieval of the actual papers, is of course, delegated to the respective publisher and/or public repositories like Google Scholar, ResearchGate,Academia, etc.

Thus, overall the plan of the study is *to perform searches to the dblp online publication repository and to download the papers that pass throught the inclusion/ exclusion filters.*. Moreover, the plan includes *the validation of the cor-*

[1] See `https://dblp.org/faq/index.html` for details

pus collected on the basis of a seed of well-known papers in the area of query optimization using machine learning techniques (which we will check on whether is part of the resulting corpus - see section 3.2.4).

3.2.3 Search Strings and Search scope

We've chosen to focus our search on the period [2018 - 2023] due to the fast-paced evolution of machine learning. This timeframe ensures the inclusion of the latest techniques in this rapidly advancing field. Additionally, given that machine learning is a relatively new branch, this approach captures contemporary contributions and reflects its dynamic nature.

Our definition of search strings is based on careful inspection of the titles and keywords of a corpus of representative papers that we have empirically selected, based on our knowledge of their frequency in the literature and their spread in time. We have excluded keywords of very broad nature and focused only on terms that are representative of the topic.

Fundamentally, the search terms are produced by the Cartesian Product of three sets of terms:

{Query}
x
{Optimization/Optimizer}
x
{Machine/Learned/ML/Learning/ Artificial Intelligence/Deep/AI}

Essentially, the search string is produced by AND-ing the three sets. Practically, instead of firing a single keyword query in disjunctive normal form at dblp, the plan is to pick all the pairs of terms consisting of the term "query" and one term from the second set and another from the third set, and firing each such pair separately. Thus, the individual queries to be fired are of the form:
("Query" AND "Optimization" AND "Machine Learning")
("Query" AND "Optimization" AND "ML")
("Query" AND "Optimization" AND "Learning")
("Query" AND "Optimization" AND "Artificial Intelligence")
("Query" AND "Optimization" AND "Deep")
...
("Query" AND "Optimizer" AND "Learned")

3.2.4 Study Validation

To assess the validity of the study, we selected the following papers as the seeds of our study: [3] , [6] , [4] , [7]

3.3 Team

?????

3.4 Study Selection

We follow the related literature on explicitly listing our inclusion and exclusion criteria. For a paper to be included, all the inclusion criteria must be met. For a paper to be excluded, a single exclusion criterion is sufficient. In the sequel, we present our inclusion and exclusion criteria, as well as the protocol for screening papers as well as for resolving ambiguities[8], [5].

Inclusion Criteria

We include publications fulfilling all of the following criteria:

(i) they use Machine Learning or AI techniques for query optimization
(ii) they were published in the time-interval 2018-2023
(iii) they have clearly been peer-reviewed
(iv) only in English language.

Exclusion Criteria

We exclude publications fulfilling all the following criteria:
(i) studies that were not authored in English
(ii) studies of reference nature , summarizing results of other publications – indicatively, but not exhaustively: Books; Chapters, or in general, parts of Books or Collections; Encyclopedia lemmas; Theses of all kinds (PhD, MSc, Diplomas or other)
(iii) non-original studies (e.g., multiple copies of the same publication)
(iv) studies not matching the inclusion criteria:
• studies not concerning the usage of ML or AI techniques for query optimization
• studies that were published outside the reference time-interval
• studies that have not been clearly peer reviewed
• studies that were published in other language than English

3.5 Data Collection and Extraction

In this subsection, we discuss, what we did in order to collect and organize the required publications. To collect the data we will proceed as follows:

1. First, we will pose the search strings to the dblp online engine, via its dedicated search facility `https://dblp.org/search/` and retrieve the resulting HTML page, the respective bibliography file (.bib) and XML files for each query.
2. Second, we will eliminate duplicate entries from the multiple searches that will be performed in the previous step.
3. Third, we will also exclude all publications that obviously do not fit our inclusion criteria. At the end of this step, we will be with a single list of "final" survivor entries.
4. Forth, we will proceed to collect the survivor publications locally, as PDF files.

3.6 Plan of Classification

In this section, we discuss the plan on how to classify the different papers with respect to the metadata we will need to assign to each of them, in order to be able to address the original research questions.

3.6.1 Addressing the research question of the annual activity - RQ1

Concerning the answering of the overall annual breakdown and trend of publication, we will need to keep the *year* of the publication, as reported by dblp. For our studies we will focus in time period 2018-2023.

3.6.2 Addressing the research question of the venue and type - RQ2

Venue Type. We originally consider the following "venue types" of publi-

cations that comply with the inclusion criteria:
• Journal Articles
• Conference and Workshop Papers

Content Type. Orthogonally to the above classification, and concerning the essence of the contribution of the papers, we will also discriminate them in terms of "essence type" in the following categories:
• Research paper
• Demo paper
• Vision or Opinion paper
• PhD Proposal
• Survey
• Tutorial
• Other

3.6.3 Addressing the research question of identifying the specific domain in which each paper discusses the utilization of machine learning techniques for query optimization - RQ3

Our research endeavors involve a comprehensive examination of each paper to ascertain the specific domain in which query optimization using only machine learning techniques is applied.

3.6.4 Addressing the research question of the topics and the methods employed of the surveyed papers - RQ4

Regarding the research inquiries posed by each paper, each paper may encapsulate one or more of the ensuing research contributions:

- Is it feasible to employ machine learning for query optimization without reliance on an established expert optimizer?
- How can query optimization performance and adaptability be advanced through the utilization of machine learning techniques?
- What strategies encompass the application of machine learning techniques to database systems, particularly emphasizing query optimization enhancements?
- In what ways can existing traditional methods for query optimization be augmented using machine learning approaches?
- How can the development of a federated query optimizer capable of seamless integration with diverse SQL-based database systems be achieved with minimal engineering effort?
- What methodologies facilitate the optimization of cloud query processing by harnessing the power of reinforcement learning techniques?
- In what ways can machine learning contribute to adaptive and self-tuning query optimization mechanisms?
- Other

Methods employed. The methods that a paper employs to address the research question can be one or more of the following method types.

- Algorithm
- Architecture of a query optimizer
- Theoretical Proof
- Other

3.6.5 Investigating the Architectural Aspects of Machine Learning-based Query Optimizers - RQ5

In addressing the research inquiry surrounding the architectural dimensions of query optimizers based on Machine Learning, our investigation focuses on comprehensively examining papers that fall within our defined inclusion criteria. Specifically, we scrutinize each paper to ascertain whether it expounds upon any form of architectural framework for the proposed query optimizer.

3.6.6 Addressing the Research Question of Machine Learning Techniques Employed - RQ6

In addressing this research question, we undertake a systematic review of each selected paper with a specific focus on identifying the machine learning techniques employed. Our evaluation criteria are broad; we welcome the inclusion of any technique that falls under the umbrella of machine learning. Our primary objective is to understand the range and applicability of machine learning techniques in the context of query optimization. By doing so, we aim to provide a comprehensive overview of the current landscape and emerging trends in machine learning-driven query optimization techniques.

3.6.7 Addressing the Research Question of Model Training - RQ7

To address this particular research question, we delve into the methodologies adopted by each paper to train their respective machine learning models. We are not anchored to any specific training model.

3.6.8 Addressing the Research Question of Source Code Availability - RQ8

To address RQ8, we meticulously review each paper to ascertain the availability of source code or detailed implementation blueprints. Our focus is on discerning the transparency and reproducibility of the proposed solutions.

3.6.9 Addressing the Research Question of Database Systems Tested - RQ9

In response to RQ9, our examination revolves around identifying the specific database systems upon which the proposed query optimization techniques have been tested or implemented. We maintain an inclusive approach; every database, irrespective of its type or popularity, is welcomed for consideration. We are not anchored to any specific systems, allowing us to appreciate the full spectrum of applications.

3.6.10 Addressing the Research Question of Search Space Pruning Methods - RQ10

For RQ10, we assess each paper for methods related to search space pruning in the realm of query optimization. We adopt an inclusive stance, welcoming any and all techniques presented. Our goal is to encapsulate the range of methods in current literature, providing a holistic view of advancements and applications in this area.

3.6.11 Addressing the Research Question of Query Optimizer Comparison - RQ11

In approaching RQ11, we closely examine each paper to discern how the proposed query optimizer is juxtaposed against existing ones.

3.6.12 Addressing the Research Question of Limitations or Drawbacks - RQ12

In addressing RQ12, our attention is directed towards pinpointing any outlined limitations or drawbacks associated with the machine learning techniques proposed for query optimization. We sift through each paper, seeking candid acknowledgments of potential shortcomings or areas of improvement. This exercise aims to offer a balanced view, highlighting both the merits and the areas of caution associated with these emerging techniques.

3.6.13 Addressing the Research Question of Further Research Avenues - RQ13

For RQ13, we navigate through the discussions and conclusions of each paper, identifying suggestions or implications for further research in the domain of query optimization. Our objective is to collate potential pathways and promising areas that could shape the next wave of advancements in this field. By doing so, we aspire to provide a roadmap for researchers and practitioners keen on exploring beyond the current horizons.

4 Execution of the protocol and corpus compilation

In this Section, we report on the steps we followed to execute the collection protocol, whose design we have reported in (Figure 2).

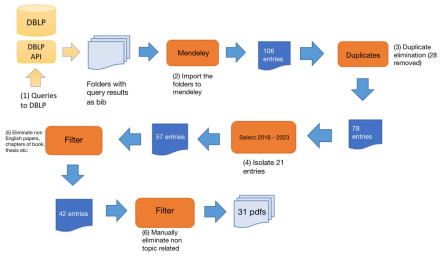

Figure 2
The process of corpus compilation

4.1 Preliminary collection of candidate references

In the previous section, we have described in detail the plan of the collection of (a) the necessary metainformation from dblp, and, (b) the corpus of publications per se from the all available sources online. We have implemented the compilation process as described in the plan.
The initial phase involved deploying our search strings to the DBLP database. To systematize the resultant outputs, the bibliography file ('.bib') of each query was procured and subsequently imported into Mendeley[2], a distinguished reference management software. Mendeley aids in the methodical collection, organization, and citation of research sources, ensuring

[2] https://www.mendeley.com/guides/desktop/

meticulous and efficient handling of the scholarly materials. The parsing produced a list of 106 publications overall.

The subsequent phase entailed processing the amassed information and vigilant scrutiny for duplicate entries. Utilizing Mendeley's robust functionalities, potential duplicates were systematically identified through its dedicated 'Check For Duplicates' feature (accessible under the 'Tools' menu). Upon meticulous elimination of these redundancies, the refined list comprised 78 unique publications, indicating the removal of 28 duplicate entries.

The tertiary phase involved chronological categorization of the publications by their year of issuance. Publications outside the timeframe of 2018 to 2023 were systematically excluded. Mendeley's intuitive interface greatly facilitated this process, offering clear visibility and quick access to the publication year of each document. Post this rigorous filtering, a corpus of 57 documents remained, resulting in the exclusion of 21 publications from the initial collection.

The fourth phase necessitated further refinement of the document collection. Publications not authored in English were systematically excluded, resulting in the omission of one such paper. Additionally, entries classified as book chapters, references to other works, or academic theses were meticulously identified and removed, leading to the exclusion of an aggregate of 14 papers. Subsequent to these evaluations, the finalized collection comprised 42 scholarly publications. It's noteworthy that Mendeley's intuitive interface was instrumental in efficiently facilitating these classification and exclusion processes.

In the concluding phase, an exhaustive manual review was undertaken to discern and exclude publications that did not intrinsically align with the application of machine learning techniques for query optimization. For a comprehensive evaluation, we sourced the PDF file of each listed publication from its respective bibliography entry. This rigorous process ensured the exclusion of papers that, while matching our search criteria in terms of keywords, deviated in their thematic essence. This meticulous filtration led to the removal of 11 such tangential papers. As a result, a refined collection of 31 publications stood out, each congruent with the pivotal theme of our research.

Overall, in the beginning we had collected 106 publications and after all this filtering the remaining were 31. In Tables 2, 3, 4,56 and Figures 3 and 4 we report the distribution of papers per year and type for the removed entries as well as the survivors of the process. In the last Table 6 we can see the survivors of the filtering process.
Taking a look in Tables 2, 3 and 4 we can see that in total, 40 entries were discarded due to either duplication or being outside the specified time frame of 2018-2023. Initially, we had identified 28 duplicates and 21 outside the period, summing to 49. The discrepancy of 9 arises from entries that were counted multiple times as duplicates.

Table 2
Breakdown of candidate reference entries removed during steps 3 (duplicate removal) and 4 (filtering for the period 2018-2023). The table has been split in three for clearer presentation for these periods [1982 - 2007], [2009 - 2019] and [2020 - 2023]

	Years								Total
	1982	1991	1992	1996	1997	2000	2006	2007	
Book and Thesis									
Conference and Workshop Papers		1	1		1		1	1	5
Journal Articles	1		1			1			3
Parts in Books				1					1
Total	1	1	2	1	1	1	1	1	9

Table 3
Breakdown of candidate reference entries removed during steps 3 (duplicate removal) and 4 (filtering for the period 2018-2023).

	Years								Total
	2009	2012	2013	2014	2015	2016	2018	2019	
Book and Thesis									
Conference and Workshop Papers	1	1	1				1	1	5
Journal Articles		1	1			2	2	4	10
Parts in Books									
Total	1	2	2			2	3	5	15

Table 4
Breakdown of candidate reference entries removed during steps 3 (duplicate removal) and 4 (filtering for the period 2018-2023).

	Years				Total
	2020	2021	2022	2023	
Book and Thesis			2		2
Conference and Workshop Papers	2		4	1	7
Journal Articles	1	1	3	2	7
Parts in Books					
Total	2	3	8	3	16

Table 5
Breakdown of candidate reference entries removed during steps 6 (english only, exclude chapters of books, thesis etc.) and step 7 (non-topic related)

	Years							Total
	2017	2018	2019	2020	2021	2022	2023	
Conference and Workshop Papers		1	1		1	4	1	8
Journal Articles	2	1	2	1	5	5	2	18
Total	2	2	3	1	6	9	3	26

Table 6
Final statistics for the papers that survived filtering.

	Years						Total
	2018	2019	2020	2021	2022	2023	
Conference and Workshop Papers	1	1	2	2	8	2	16
Journal Articles	1	2	2	1	3	6	15
Total	2	3	4	3	11	8	31

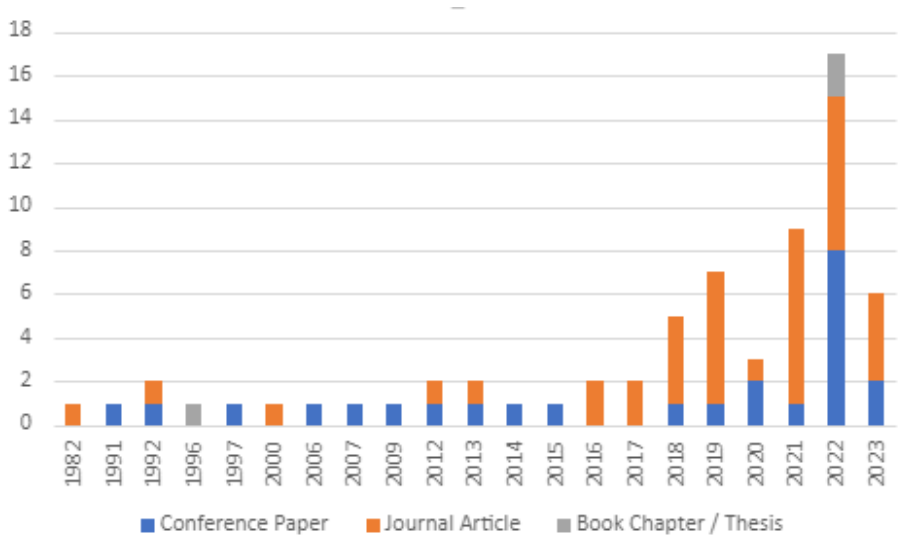

Figure 3
Bar chart of entries that were removed via step 3 (duplicate elimination), step 4 (Out of date), step 5(non English papers, chapters of book, thesis, etc.) and step 6 (manually elimination of non topic related papers). The categories listed in the figure include all possible categories, independently of inclusion or exclusion criteria.

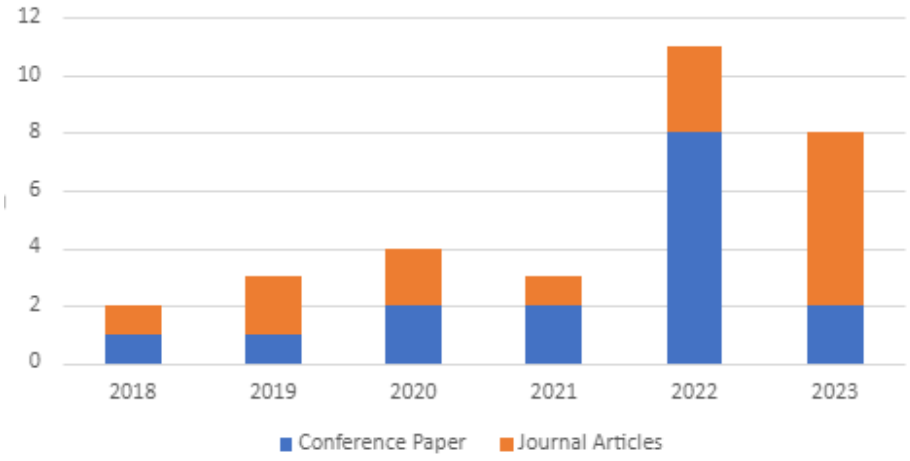

Figure 4
Bar chart of the 31 papers that passed all filtering criteria and met the inclusion requirements.

5 Results

Once the papers were collected, we proceeded in answering the research questions of our plan. In this section, we discus the findings of our analysis, organized by research questions.

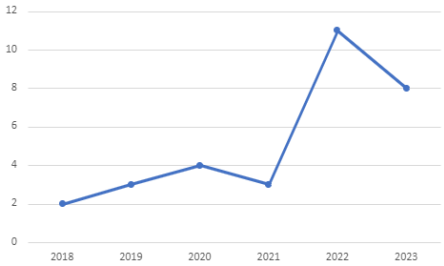

Figure 5
Annual number of publications for the period [2018 - 2023]

5.1 RQ1 Results: What is the annual amount of publications and the overall trend?

The first research question was concerned with answering "What is the annual amount of publications and the overall trend?". Figure 5 depicts the results.

Practically, we observe a discernible growth trend. Beginning with 2 publications in 2018, there's a consistent rise, peaking at 11 in 2022. However, 2023 shows a mild dip with 8 publications. Of particular interest is the substantial increase in 2022, followed by a decrease in 2023. Nonetheless, it's crucial to note that 2023 is not yet over, suggesting the possibility of more publications being on the horizon. The

data overall hints at periods of growth, potential shifts in research interest or activity, and the evolving dynamics of academic output over these years.

5.2 RQ2 Results: What is the breakdown of publications in terms of venues and types?

The second research question has to do with "What is the breakdown of publications in terms of venues and types?". Expectedly, the majority of the papers, exceeding 51.6% are conference papers (Figure 6). Similarly, as we can see in Figure 7 most of the paper's content type is conference paper.

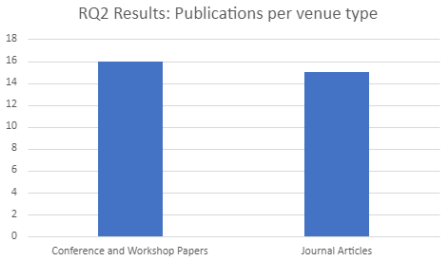

Figure 6
Breakdown of publications per venue

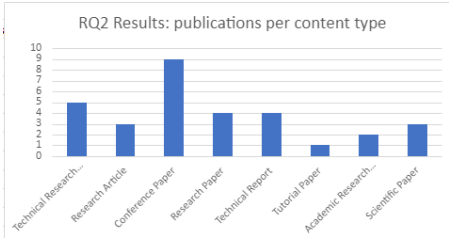

Figure 7
Breakdown of publications per content type

	Conference & Workshop Papers	Journal Articles	Total
2018	1	1	2
2019	1	2	3
2020	2	2	4
2021	2	1	3
2022	8	3	11
2023	2	6	8
Total	16	15	31

Figure 8
Annual breakdown of publications per venue

Figure 8 illustrates the yearly distribution of publications by venue type. While the number of conference papers remains consistent over the years, barring a spike in 2022, journal articles exhibit stability until a noticeable uptick in 2023.

Figure 9 provides an annual breakdown on research effort by the content type of the publications. All kind of content type have a fairly stable rate, with the exception of the spike of 2022.

	2018	2019	2020	2021	2022	2023	Total
Conference & Workshop Papers	1	1	2	2	8	2	16
SIGMOD	1	1			2		4
VLDP					1		1
CIDR					2		2
MEDI				1			1
BTW						2	2
EDBT					1		1
CODS				1			1
ICDE			1				1
aiDM			1				1
DaWak					1		1
GIS					1		1
Journal Articles	1	2	2	1	3	6	15
J. Electr. Comput. Eng.	1						1
VLDP		1		1	1	3	6
IEEE Access					2		2
CoRR		1	1			3	5
SIGMOD Rec.			1				1
Total	2	3	4	3	11	8	31

Figure 10
Annual breakdown of publications per venue

	Technical Research Paper	Research Article	Conference Paper	Research Paper	Technical Report	Tutorial Paper	Academic Research Article	Scientific Paper	Total
2018	2								2
2019			1	1			1		3
2020	1			1			1	1	4
2021		1	1		1				3
2022		1	6	1	1	1		1	11
2023	2	1	1	1	2			1	8
Total	**4**	**3**	**9**	**4**	**4**	**1**	**2**	**3**	**31**

Figure 9
Annual breakdown of publications per content type

In Figure 10, it is evident that SIGMOD stands out as the primary venue for research on machine learning techniques in query optimization. CIDR and BTW also contribute, each hosting several papers. The breadth of venues featuring such research indicates an emerging trend in the field. Among journals, VLDB leads in terms of hosting the most publications, followed by CoRR.

5.3 RQ3 Results: In which specific field does this paper discuss the application of machine learning for query optimization?

The third research question involves answering the question "In which specific field does this paper discuss the application of machine learning for query optimization?".

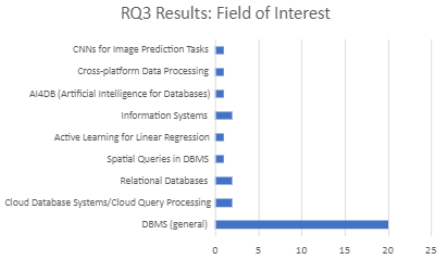

Figure 11
Breakdown of fields that discuss the application of machine learning for query optimization.

In Figure 11, it's evident that the predominant focus revolves around the broad realm of database management systems. However, this figure doesn't delve into the nuanced specifics within this domain. Some other fields are Cloud Database Systems, Relational Databases and Information Systems.

5.4 RQ4 What are the primary research questions and solution methods proposed in this paper concerning query optimization using machine learning?

The forth question involves the answering of two problems: "what are the research questions" and "what are the types of methods used to answer the research questions".

In Figures 12 and 13, we address the aforementioned research questions. From Figure 12, it is evident that 41 papers span the various topics, with some papers delving into multiple research questions. Conversely, Figure

13 encompasses 33 papers, reflecting instances where papers adopt both the Architecture of a Query Optimizer approach and alternative methods, such as Theoretical Proof.

On a broader perspective, there is a pronounced inclination towards enhancing query optimizers using Machine Learning (ML) techniques and addressing the inherent challenges posed by the integration of ML in query optimization. Notably, leveraging architectures for ML-driven query optimization emerges as the prevailing methodology in the literature.

In Figure 14 we present how the topics have evolved over time (observe that the grand total is not 31, but 41, due to the presence of more than one topics in some papers). The topics of the study of how to enhance existing query optimizers using Machine Learning (ML) techniques and addressing the inherent challenges posed by the integration of ML in query optimization, seem to be constantly present in the agenda and to retain their dynamics. The rest of the topics seem to be low in their presence in the literature.

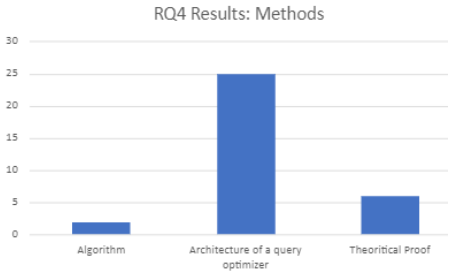

Figure 13
Breakdown of publications per method used to answer the research questions

5.5 RQ5 Does the paper present a specific architecture for a machine-learning-based query optimizer, or does it theorize on optimal structures without detailing a concrete architecture?

The fifth question was "Does the paper present a specific architecture for a machine-learning-based query optimizer, or does it theorize on optimal structures without detailing a concrete architecture?".

In Figure 15 we can see that from 2018 to 2023, there has been a noticeable upward trend in the number of publications focusing on the use of architecture for query optimization with machine learning. Starting with a solitary publication in 2018, the interest in this area has grown substantially, peaking in 2022 with seven publications. This suggests a rising emphasis and recognition of the potential of integrating architecture with machine learning techniques for query optimization. The slight dip in 2023, with six publications, may indicate a consolidation phase or a shift in research focus, but overall, the heightened interest in this domain over the years is unmistakable.

Out of a total of 31 papers that were scrutinized, 25 publications specifically delve into the architecture of a machine learning query optimizer. This leaves a subset of six papers that veer away from architectural descriptions; instead, these primarily hinge on theoretical proofs or algorithmic

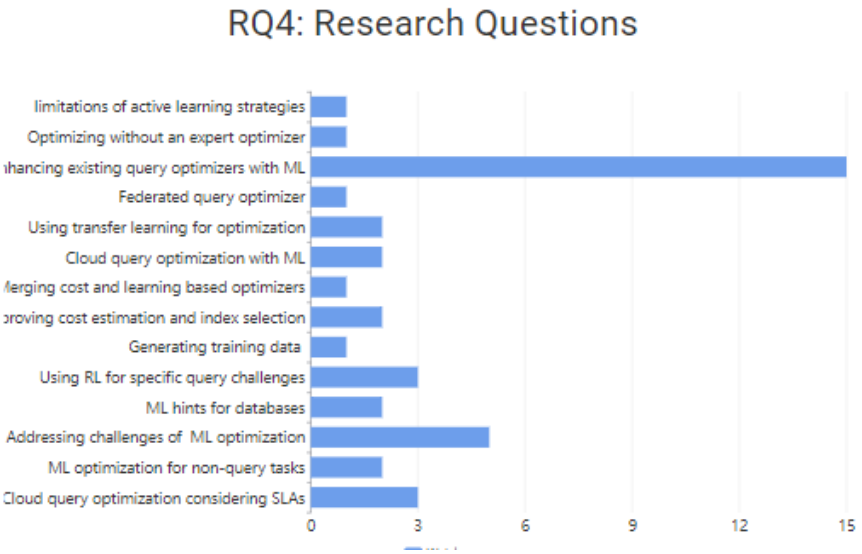

Figure 12
Breakdown of publications per research question

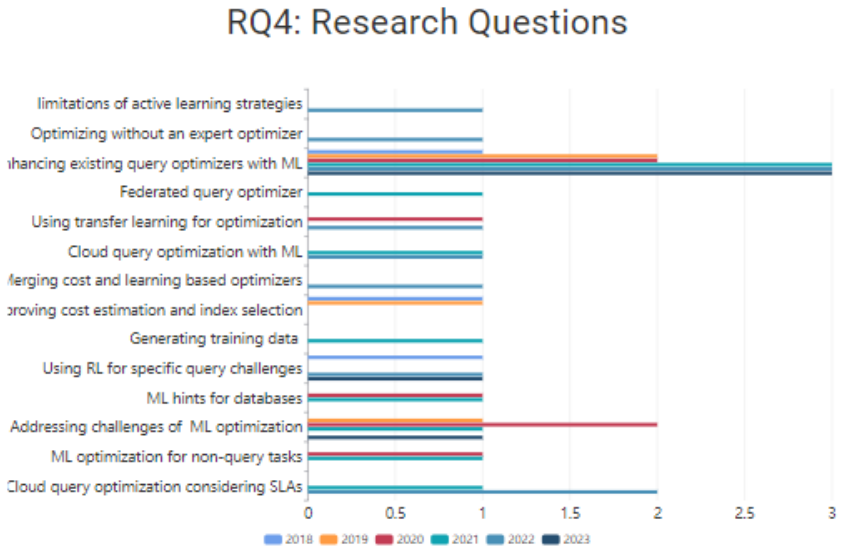

Figure 14
Breakdown of publications per research question

methodologies. It underscores the predominance of architectural insights

in the wider discourse, while also highlighting the alternative approaches embraced by a fraction of the studies.

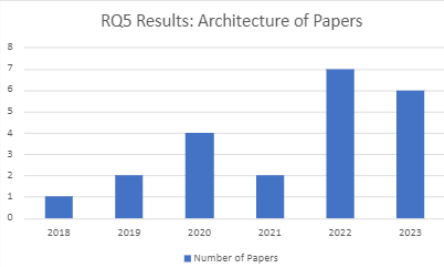

Figure 15
Annual breakdown of publications that use architecture or other methods

5.6 RQ6 What machine learning technique does the paper employ for query optimization, and what outcomes are achieved through its implementation?

The sixth query posed was, "Which machine learning technique is utilized for query optimization in the paper, and what results stem from its application?". To provide a comprehensive response, we deconstructed this question into two distinct inquiries: "Which machine learning techniques are employed in each paper?" and "What specific outcomes result from the implementation of each ML technique?".

In Figure 16, the distribution of publications by Machine Learning Technique is presented. Deep Reinforcement Learning emerges as the predominant technique, followed closely by (Deep) Neural Networks. These methods appear to yield the most encouraging outcomes.

Figure 17 offers an annual distribution of publications based on the Machine Learning Technique from 2018 to 2023. Once again, Deep Reinforcement Learning and Deep Neural Networks consistently maintain their prominence throughout the years.

It's noteworthy that while the total number of publications across Figures 16 and 17 is 26, there are 31 papers in all. The discrepancy arises as the remaining 5 papers deploy generic Machine Learning techniques without honing in on any specific approach.
We now turn our attention to the advancements achieved through the application of Machine Learning techniques in query optimization. The key accomplishments include:

- Elevated performance levels in regression models.
- Cost-effective data annotation processes.
- Rapid acceleration in optimization tasks.
- Efficacious utilization of limited datasets.
- Notable improvements in query execution efficiency and cost metrics.
- Marked superiority over conventional query optimization strategies.
- Precision in both selectivity estimation and query performance predictions.
- Enriched user experience through selectable optimization hints.
- Demonstrable resilience and adaptability in fluctuating

environments.

- Exactness in estimating query cardinality.
- Streamlined generation and execution of query plans.
- Enhanced capability to recognize patterns and predict outcomes.
- A distinct edge in performance when compared to traditional methodologies.

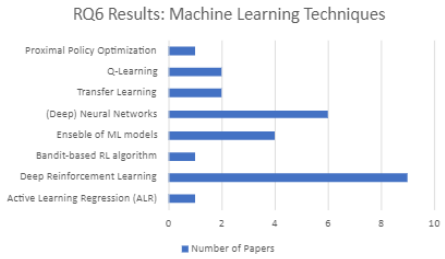

Figure 16
Breakdown of different Machine Learning techniques

5.7 RQ7 How is the training of this machine learning model accomplished in the study?

The seventh question concerns with "How is the training of this machine learning model accomplished in the study?".

Figure 18 showcases the variety of training methods found in the publications and how often each method is used. On the other hand, Figure 19 offers a year-by-year look at the popularity of these methods, revealing trends and shifts over time. Together, these figures give a clear picture of the landscape of machine learning training methods in our study's scope.

From Figure 18, it's evident that (Deep) Reinforcement Learning (DRL) is the leading training method in the papers we examined. The prominence of DRL in training models stems from its ability to let models learn optimal strategies from trial and error. Unlike traditional methods, DRL continuously fine-tunes the model's decisions based on rewards, making it particularly potent for training models that need to adapt to changing conditions or learn complex, non-linear strategies.

Supervised learning, the second most prevalent method, is favored for its structured approach. When training models, supervised learning provides a clear framework where models are nurtured using labeled datasets. This clear guidance ensures that models have direct feedback on their predictions, making it an excellent choice for training models where there's an abundance of historical data with known outcomes.

In contrast, while DRL and supervised learning are predominant, other methods find less representation in the literature. Their limited presence could suggest niche applications or emerging methods still solidifying their footing in the realm of model training.

	2018	2019	2020	2021	2022	2023	Total
Active Learning Regression (ALR)					1		1
Deep Reinforcement Learning	1	2	1		3	2	9
Bandit-based RL algorithm			1				1
Enseble of ML models					3	1	4
(Deep) Neural Networks	1		1	1	1	2	6
Transfer Learning			1	1			2
Q-Learning		1		1			2
Proximal Policy Optimization					2		1

Figure 17
Annual breakdown of different Machine Learning techniques

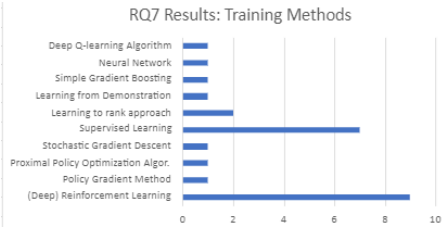

Figure 18
Breakdown of different training techniques

From Figure 19, the annual dispersion of publications based on training methods is showcased. Notably, both (Deep) Reinforcement Learning (DRL) and supervised learning have maintained a consistent presence over the years.
While our study encompasses a total of 31 papers, only 25 explicitly specify their training methods. This disparity can be attributed to a subset of papers that either omit mention of a distinct training method or refer to broad machine learning methodologies without delving into specific techniques. It's also noteworthy that some papers are multifaceted, incorporating more than one training method to harness multiple techniques in their approach. This suggests a dynamic field where researchers are exploring combinations to optimize model training.

5.8 RQ8 Does the paper offer any source code or implementation details for the proposed query optimizer?

The eights question concerns "Does the paper offer any source code or implementation details for the proposed query optimizer?".

From Figure 20, the annual distribution of publications offering source code is delineated. Distinctly, out of the 31 examined papers, a mere 8 provide some form of code accompanying their optimizer. This limited inclusion of code is rather disheartening. The lack of accessible code can hinder replication studies, stifle further advancements in the field, and create barriers to real-world application. Source code can act as a foundational block, allowing other researchers to build upon previous works, ensuring a richer, more col-

	2018	2019	2020	2021	2022	2023	Total
(Deep) Reinforcement Learning		2		2	2	3	9
Policy Gradient Method			1				1
Proximal Policy Optimization Algor.					1		1
Stochastic Gradient Descent	1						1
Supervised Learning		1	1	1	3	1	7
Learning to rank approach						2	2
Learning from Demonstration		1					1
Simple Gradient Boosting						1	1
Neural Network			1				1
Deep Q-learning Algorithm					1		1
Total	**1**	**4**	**3**	**3**	**7**	**7**	**25**

Figure 19
Annual breakdown of different training techniques

laborative, and progressive research environment.

Upon closer observation of Figure 20, we note a yearly trend: typically, only one publication annually shares code. However, the years 2019 and 2023 stand out, with two and three papers respectively providing code. This could signal an emergent shift in the community, possibly indicating a growing recognition of the importance of transparency and reproducibility. As such, one could optimistically infer that upcoming studies may increasingly embrace the practice of sharing code, fostering a more inclusive and robust research ecosystem.

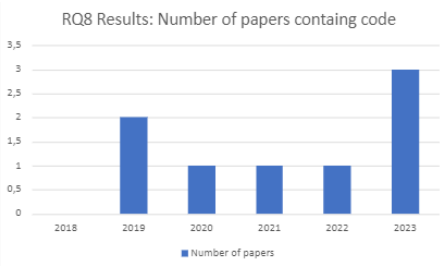

Figure 20
Annual breakdown of access in source code for the optimizers.

5.9 RQ9 On which database systems has the optimization been tested or implemented?

The ninth question concerns "On which database systems has the optimization been tested or implemented?".

In dissecting the question concerning database types utilized in the

research, two distinct figures provide clarity. From Figure 21, we derive insights into the varied databases and their respective adoption across publications.

A striking observation is that a predominant chunk of papers, precisely 8, remain ambiguous, employing a medley of databases without zeroing in on a specific one. On the other hand, PostgreSQL stands out as a favored choice, with 6 studies leveraging it for their machine learning optimizer endeavors. The remaining databases feature sporadically in the literature, signaling their niche or specialized appeal.

Of the whole corpus, only 18 papers lay clear emphasis on their database of choice, rendering them explicit in their methodologies. Conversely, the absence of a specified database in the residual papers is a notable gap. Such omissions can potentially cloud replicability, limiting the understanding of the context in which the proposed solutions or methods are most effective.

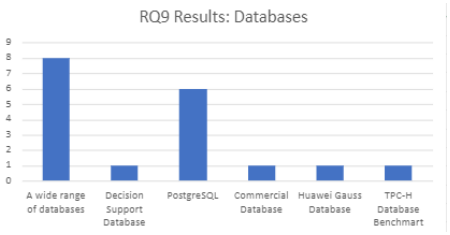

Figure 21
Breakdown of databases

In Figure 22, the chronological progression of database choices across publications is illuminated. Year-by-year, a recurring theme emerges: a significant portion of research does not anchor itself to any specific database. This widespread nondescript choice hints at the versatility of the models proposed, suggesting their potential adaptability across various database platforms. Such adaptability is a commendable attribute, as it implies broader applicability in diverse real-world scenarios.

Yet, amidst this broad spectrum, PostgreSQL consistently garners attention, asserting its relevance and reliability in the field over time. Its recurrent appearance across the years underpins its robustness and perhaps its alignment with the evolving demands of machine learning optimizers.

Conversely, the sporadic representation of other databases in the timeline indicates their selective or situational utility in the research realm. For scholars and practitioners alike, this mapping study underscores the predominant trends, aiding in informed choices for future endeavors in the dynamic landscape of machine learning optimizers for databases.

	2018	2019	2020	2021	2022	2023	Total
A wide range of databases	2	1	2		1	2	8
Decision Support Database				1			1
PostgreSQL			1		2	3	6
Commercial Database						1	1
Huawei Gauss Database					1		1
TPC-H Database Benchmart						1	1
Total	2	1	3	1	4	7	18

Figure 22
Annual breakdown of databases

5.10 RQ10 Which methods for search space pruning are highlighted or employed in this study?

The tenth question concerns with "Which methods for search space pruning are highlighted or employed in this study?".

Figure 23 offers an intriguing dissection of the pruning methods embraced by publications. At a glance, diversity reigns, with each method predominantly making a singular appearance across the surveyed literature. Only "coarse-grained hints" deviate from this trend, marking its presence twice. This might indicate a budding preference or its efficacy in the context.

However, when we unpack the yearly distribution in Figure 24, 2023 stands out as a year of heightened exploration in pruning methods for machine learning optimizers. This could be a harbinger of an emerging research direction, suggesting that subsequent studies may delve deeper into refining and employing pruning techniques.

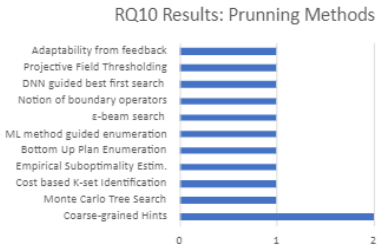

Figure 23
Breakdown of pruning methods

	2018	2019	2020	2021	2022	2023	Total
Coarse-grained Hints			1			1	2
Monte Carlo Tree Search					1		1
Cost based K-set Identification						1	1
Empirical Suboptimality Estim.						1	1
Bottom Up Plan Enumeration						1	1
ML method guided enumeration						1	1
ε-beam search					1		1
Notion of boundary operators			1				1
DNN guided best first search		1					1
Projective Field Thresholding			1				1
Adaptability from feedback		1					1
Total		2	3		2	5	12

Figure 24
Annual breakdown of pruning methods

Out of 31 papers, the fact that only 12 detail their pruning method is noteworthy. Integrating pruning methods in machine learning optimizers for query optimization is paramount because it aids in reducing the search space, thus optimizing the efficiency of query processing. By narrowing down potential solutions, pruning can expedite the optimizer's decision-making, potentially leading to faster and more resource-efficient query results.

5.11 RQ11 How does the query optimizer proposed in this paper compare with existing query optimizers in terms of databases used and overall performance?

The eleventh question concerns with "How does the query optimizer proposed in this paper compare with existing query optimizers in terms of databases used and overall performance?".

Figure 25 shows the optimizers used in the studies. Figure 26 details the comparisons made between them

(already existing optimizers and the brand new ones). Figure 27 displays the year-by-year trends of these comparisons. Together, they give a clear picture of which optimizers are popular and how they're compared over time.

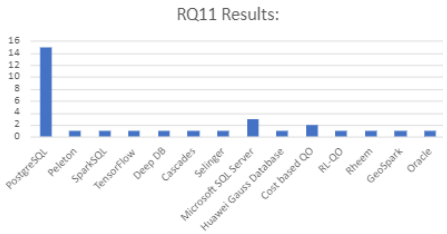

Figure 25
Breakdown of optimizers

Figure 25 displays the various optimizers that studies have benchmarked against. Notably, PostgreSQL stands out as the most frequently compared optimizer. This prevalence might be attributed to PostgreSQL's widespread use, open-source nature, and its reputation for robust performance. Additionally, it offers a benchmark for new techniques, given its established and well-documented query optimization strategies. The chart also highlights a diverse set of other optimizers used for comparison in the research.

Figure 26 presents a compelling visual comparison of newer optimizer techniques against established ones. On the horizontal axis, we observe the innovative optimizers introduced by recent studies. Conversely, the vertical axis showcases widely-recognized existing optimizers. The plotted dots signify instances where the new optimizer techniques surpass their traditional counterparts. This dominant performance of emerging optimizers suggests a promising shift, potentially heralding a transformative phase in optimization strategies.

In Figure 27, we're presented with a year-by-year dissection of optimizers employed in papers for comparative analyses. Notably, PostgreSQL and Microsoft SQL Server consistently appear throughout the years, suggesting their enduring relevance and perhaps their benchmark status in optimization studies. It's interesting to note that the total count of referenced optimizers stands at 36, surpassing the 31 papers in this study. This discrepancy underscores that several papers have chosen to employ multiple optimizers for a more comprehensive comparative evaluation, enriching the depth of their findings.

	2018	2019	2020	2021	2022	2023	Total
PostgreSQL		1	2	2	6	6	17
Peleton					1	1	2
SparkSQL					1		1
TensorFlow					1		1
Deep DB					1		1
Cascades				1	1		2
Selinger				1			1
Microsoft SQL Server		1	1			2	4
Huawei Gauss Database					1		1
Cost based QO						2	2
RL-QO					1		1
Rheem			1				1
GeoSpark				1			1
Oracle		1					1
Total		3	4	5	13	11	36

Figure 27
Annual breakdown of optimizers

Figure 26
Comparison of optimizers. The horizontal axis is the new proposed optimizers and the other axis is the already existing optimizers/

5.12 RQ12 What are the identified limitations or drawbacks of the machine learning technique proposed in this study for query optimization?

In our penultimate research query, we delve into a critical examination: "What limitations or challenges are inherent to the machine learning techniques proposed for query optimization?".

Across the 31 papers studied, it's evident that like all tools, machine learning isn't devoid of pitfalls. However, it's crucial to view these in the broader context: while there are challenges, machine learning's strengths are undeniable.

Recall our prior analyses, where we drew comparisons between emerging machine learning techniques and established optimizers like PostgreSQL and Microsoft SQL Server. These side-by-side evaluations underscored machine learning's unique ability to tackle longstanding issues in cost-based query optimization. Traditional models often grapple with striking the right balance between fine-tuning query plans and the associated computational costs. Additionally, they can be challenged by the unpredictable nature of real-world queries, a domain where machine learning techniques often demonstrate superior adaptability.

With that perspective, let's unpack the limitations while keeping an eye on the broader horizon of machine learning's potential:

1. **Data Dependency**: Machine learning models are intrinsically tied to the quality and quantity of training data, making the acquisition of accurate, representative query plans a challenge, and necessitating frequent retraining with updated datasets.
2. **Model Interpretability**: Particularly in models like Deep Reinforcement Learning (DRL), there's an inherent complexity in understanding their decision-making, underscoring the need for transparency and insights to ensure trust in ML-based solutions.
3. **Computational Overheads**: The computational demands of training complex models, like DRL, are significant. This is exacerbated by the associated costs of collecting training data and

the necessity for exhaustive hyperparameter tuning during optimization.

4. **Generalization Concerns**: Models designed for specific datasets can falter with unfamiliar data. Moreover, adapted models might not be naturally optimized for new database environments.
5. **Model Sensitivity**: The performance optimization of many ML models, notably DRL, can be highly sensitive to hyperparameter choices, leading to variability in outcomes.
6. **Feature Extraction Challenges**: Extracting pivotal features is not straightforward, often requiring expert knowledge. The effectiveness of the model can hinge on the chosen features.
7. **Integration Difficulties**: Melding ML techniques with traditional database systems can be technically intricate. The unpredictable nature of machine learning might also misalign with the consistent expectations of conventional database systems.
8. **Training Data Challenges**: Procuring comprehensive and accurate training data is a significant obstacle, as is obtaining the valuable ground-truth labels essential for effective model training.

Despite these limitations, it's essential to perceive them not as roadblocks, but as challenges poised for innovative solutions. The current efficacy of these models underscores their potential, but acknowledging these areas of improvement can pave the way for even more refined and effective machine learning techniques for query optimization in the future. After all, every challenge presents an opportunity for advancement, and in this context, the horizon looks promising.

5.13 RQ13 Based on the findings and discussions in the paper, what avenues for further research are suggested or appear promising?

We will close our results with the question "Based on the findings and discussions in the paper, what avenues for further research are suggested or appear promising?".

Identifying future research directions is crucial for the growth of machine learning in query optimization. By pinpointing these areas, we can address current challenges and push the field forward. This approach encourages efficient use of resources, promotes collaboration among researchers, and paves the way for more advanced and effective machine learning-driven query optimizers.

1. **The Dawn of Machine Learning in Query Optimization:** Harnessing the uncharted depths of neural architectures, like RNNs and transformers, the world stands at the brink of a revolution. By blending reinforcement learning with traditional paradigms, we unveil a tapestry of possibilities. The realms of transfer learning beckon, promising a seamless amalgamation of knowledge

across tasks. We envision a future where ML techniques, fine-tuned for query optimization, orchestrate a symphony of efficiency and precision.

2. **Illuminating the Black Box - Interpretability and Integration:** As we step into a new era, the clarion call is for transparency. It's not just about having an efficient optimizer, but one whose choices are clear and discernible. How might we design components that not only perform but also communicate their rationale? In the intertwining of algorithms with system architecture, lies a path to unparalleled performance. Yet, the road to real-world deployment demands an acute understanding of challenges, blending theory with pragmatic realities.
3. **A Panoramic Vision - Broadening Horizons:** Query optimization is but a piece of the vast mosaic of database management. By viewing it as a collective entity, we find interconnections, weaving together distinct areas under the AI for Databases umbrella. Distributed databases, with their intricate symphonies, await innovative optimization techniques. We stand at a frontier, ready to extend our frameworks to realms like indexing and data cleaning, expanding the canvas of innovation.
4. **Refining the Edges - Evaluation and Performance:** Our quest for excellence doesn't end with creation; it extends to rigorous evaluation. A future awaits where models, enriched with intricate features, undergo the crucible of diverse workloads and systems. With a focus on resilience and adaptability, we envision models that not only predict but evolve. It's a journey of continuous refinement, where every iteration brings us closer to perfection.

As we embark on this journey, let us be inspired by the vastness of what lies ahead. Each avenue is not just a direction but an opportunity, a challenge, a canvas waiting for the brushstrokes of the next innovator. The future is not just about optimizing queries; it's about reshaping the very way we think about databases. For the budding researcher, the question isn't "Where can we go?" but rather, "Where do we want to take this field next?".

6 Threads to Validity

In this section, we group considerations about the validity of our design choices, our method and our results.

6.1 Validity of Explanations

Internal Validity. Internal validity refers to the credibility of providing cause-and-effect explanations between observes measurements, or, even theories and process models on how things work. In a mapping study, the goal is to provide a broader view of the "research terrain" and not to delve into discovering the consensus or debate around the truth of a research hypothesis. We are very careful to avoid any sort of explanatory comments or inferences, but strictly stick to the observed facts.

Conclusion Validity. Conclusion validity is a term that reference to the statistical significance of any observations. To the extent that there is no hypothesis testing of any kind, threats to conclusion validity are not really perceivable to our setting.

6.2 Validity of Generalizations

External Validity. External Validity refers to the ability to generalize findings from a sample to a broader population, or to a different context. We have been very specific in selecting metadata entries on the topic of query optimization using machine learning from 2018 to 2023. We are very careful to stress that (a) we do not sample in this study, but exhaustively handle all collected metadata entries and obtained document, and (b) we do not generalize our finding to other time periods or other problems (no matter how similar they might be).

6.3 Validity of the Method Used

Construct validity. Construct validity refers to the extent that the measures being studied are accurate representations of the concepts appearing in the research questions. We are explicit about our research questions, as well as the sources for our metadata entries and the inclusion and exclusion criteria. Thus, it is important to regards all our answers within the context of the aforementioned inclusion and exclusion criteria as well as the information that DBLP provides. Changes to the data source or the filtering criteria would possibly provide a different view of the research terrain.

Validation. To make sure that our metadata collection via DBLP is correct, we made sure that all the papers of the seed corpus were identified. In fact, not only were all the papers identified, but we also had a 100 percent recall in retrieving the PDF file of all of them, in the download phase (this includes both paper before and after 2018).

Search sources and terms. Probably the hottest question that one has to face in this study concerns the degree of coverage that our search achieves. In other words: is the search over DBLP, with this particular keywords, enough to guarantee an adequate map of the research terrain of query optimization using machine learning? Our answer to this -very important- threat to validity is structured along the following lines:

1. In our opinion, DBLP as a search database, is as adequate as it can get for the context of "peer-reviewed" publications, that we are interested in. To the best of our knowledge, there are no major venues of the likes of ACM, IEEE, Springer, Elsevier, and other major editors of peer-reviewed publications that are not covered by DBLP. For us, the typical motto "whatever counts is already in DBLP" is correct. Moreover, we have employed a collection of seed publications (with a significant spa thought the entire decade that we search, and with a plurality in terms of venues and authors) which was retrieved 100 percent by our queries. Having said that, it is

absolutely clear that there will be publications outside the "radar" of DBLP that pertrain to the field, and this has to be mentioned in the discussion of the threats to validity.

2. In terms of the search terms used, is it indeed possible that certain titles of paper could be outside the scope of our keywords. We intentionally expanded the scope to include other terms like "learning", "optimizer","AI" in order to broaden the scope of what is collected. Clearly, this cannot guarantee that there are indeed relevant papers that are not part of the search results. However, to the extent that all the seed corpus was returned as part of the query answer, we believe that the query space that we formulated works adequately well.
3. In term of whatever our downloaded PDF articles are the correct ones, we can say that, for all the articles for which we report results, i.e., the corpus of 31 curated papers between 2018 and 2023, the PDF was the correct one.
4. **Exclusion and inclusion of papers.** We were specific in the exclusion and inclusion list of our paper. We have collected a lot entries than actually relevant to the topic, (a) exactly because the term "optimization" is central to the computer science and (b) because the usage of the term "machine learning" in computer science is also very extensive - and DBLP is wide enough to also cover venues that pertrain in computer science. We have sampled extensively the titles of our entries, but despite this precaution measure, and despite the fact that we believe that the exclusion process was fairly successful, is is still a theoretical possibility that publications that pertain to the topic were unfairly excluded.
5. **Labeling.** Despite all the precaution measures that we have taken, is is always possible that some papers are mislabeled. However, to alleviate -or at least minimize- the possibility of error in labeling the papers with respect to the answers to research questions, we took the following precautions: (a) we went on to have an independent labeling as well as to judge discrepancies, (b) we allowed the labeling of a paper with multiple categories whatever this made sense, and, (c) we adopted a quite refined set of categories for the topics (which is obviously the most vulnerable of the research questions in terms of erroneous labeling), in order to avoid mislabelings.
6. **Researcher bias.** We had to explicitly tackle the problem of researcher bias. The avoid the intentional exclusion of material from we rely solely on the entries from DBLP. To minimize errors in the characterization of papers, we adopted a double-annotation scheme and a refereed resolution of conflicts. We do not recognize, however, that there is an imbalance due to the professor-student relationship. It is also possible that the manual removal of irrelevant papers is subject to re-

searcher bias.

7. **Papers not published.** One of the problems typically reported in the discussion of systematic reviews and systematic mapping studies is the problem of results not reported in the literature, due to a bias towards topics or opinion that are more "popular" or "trendy". Clearly, we do not address this concern: we have focused solely to papers published in conferences and journals. It is possible that the Web contains valuable material on the topic that is simply not published via the typical academic process.
8. **Repeatability.** One issue with repeatability is the extent to which the entire process is adequately documented. We believe the current description of the process gives a clear view of both the criteria used and the steps taken towards achieving the final results.

All the material of the study is available at: `https://rb.gy/2mynj`

7 Discussion

In this paper, we have performed a mapping study for the area of query optimization using Machine Learning techniques. We have designed a protocol for the mapping study, with rules for the collection of the bibliography and specific research questions to answer. According to this protocol, we have performed a search on the contents of DBLP, with well specified inclusion and exclusion criteria and collected a manually curated corpus of related literature.

Findings. We have also provided answers for several research questions that we had originally prescribed in our analysis plan. Table 7 summarizes our findings.

Usage. Coming to an end, we would also like to highlight some potential usages of this report from its readership.

The current map is not particularly suited for non-specialists, as it does not go into a layman's discussion of the area of query optimization using machine learning techniques. Clearly, the audience that we mostly address concerns researchers in the fields of data management, artificial intelligence, query optimization. We offer a map of the papers in the latest years (of course, within the scope and boundary of our inclusion/exclusion criteria). The map is, by nature, a high-level one, without going into the deep details of how different papers address the topics. Even at this high level, thought, the map of this report provides (a) the corpus of related research in one place, and, (b) the general questions of the research community during the years that are covered by the study. This can prove valuable for (a) researchers who want to survey the related literature for a specific topic, (b) reviewers who want to ensure the validity of a paper they review, (c) PhD candidates and their supervisors working towards constructing a road-map of research topics in their Thesis.

At the lower level, the knowledge

of the most studied topics and the topics with constant presence over the years can provide more details to the newcomers on where there is room for novelty, or, on the contrary, standard audience for a specific research topic. Newcomers might also benefit from knowing the venues that host publications in the area.

Finally, Editors-in-Chief and Editorial Boards of journals, as well as Program Committees of conferences can potentially see the topics beings researched and consider adding machine learning techniques for query optimization to the list of topics covered by their venue.

Table 7
Research Questions

RQ	Research Question and Answers
1	**What is the annual amount of publications and the overall trend?** The rate of publications was petty stable till 2021, with an average of 3 papers annually, but increasing in 2022 with 11 papers and in 2023 8 papers.
2	**What is the breakdown of the publications in terms of venues and types?** More than 51 percent of the papers were conference papers; also more than 48 percent were research papers.
3	**In which domain or application area does the paper investigate the utilization of machine learning techniques for query optimization?** More than 64 percent of the papers studied explore the utilization of machine learning techniques for query optimization within the general framework of database management systems.
4	**What are the primary research questions and solution methods proposed in this paper concerning query optimization using machine learning?**The exploration of enhancing traditional query optimizers with machine learning, addressing the primary challenges of machine learning techniques, and delving into the intricacies of reinforcement learning emerge as prevailing research themes. Furthermore, the adoption of novel architectures for query optimizers underscores a prevailing trend, emphasizing the pivotal role of integrating machine learning techniques into query optimization.
5	**Does the paper present a specific architecture for a machine-learning-based query optimizer, or does it theorize on optimal structures without detailing a concrete architecture?**Out of the 31 papers studied, approximately 71 percent explore the use of a novel architecture for query optimization
6	**What machine learning technique does the paper employ for query optimization, and what outcomes are achieved through its implementation?**Among the 31 papers examined, Deep Reinforcement Learning is the predominant technique, being utilized in roughly 29 percent of the papers. Additionally, Deep Neural Networks are employed in approximately 19 percent of the studies.
7	**How is the training of this machine learning model accomplished in the study?** More than 29 percent of the papers use deep reinforcement learning for training their machine learning optimizer and 23 percent of the papers use supervised learning. These stand out as the most common training methodologies in the current landscape.
8	**Does the paper offer any source code or implementation details for the proposed query optimizer?**Approximately 26 percent of the 31 papers reviewed provide some form of source code for their machine learning optimizer

9	**On which database systems has the optimization been tested or implemented?** Of the 31 papers analyzed, around 26 percent employed a diverse array of databases, while roughly 19 percent specifically utilized PostgreSQL. These statistics indicate that PostgreSQL, followed by a mix of various databases, are the prevalent systems on which machine learning query optimizers are tested or implemented.
10	**Which methods for search space pruning are highlighted or employed in this study?**A diverse array of pruning methods have been explored, with notable techniques including coarse-grained hints, Monte Carlo tree search, and cost-based K-set identification, among others.
11	**How does the query optimizer proposed in this paper compare with existing query optimizers in terms of databases used and overall performance?**Several established cost-based optimizers, such as PostgreSQL, Peleton, Microsoft SQL Server, and Rheem, have served as benchmarks for comparison with the emerging machine learning-based query optimizers. Impressively, all machine learning-enhanced query optimizers demonstrated superior performance over their cost-based counterparts.
12	**What are the identified limitations or drawbacks of the machine learning technique proposed in this study for query optimization?**Machine learning techniques for query optimization face challenges including dependency on high-quality training data, concerns over model interpretability, and computational overheads. Additionally, these models sometimes struggle with generalization across diverse data sets and environments. Integrating ML solutions into traditional database systems also presents technical hurdles.
13	**Based on the findings and discussions in the paper, what avenues for further research are suggested or appear promising?**Emerging research avenues in machine learning for query optimization encompass exploring advanced neural architectures, emphasizing model transparency and integration, and broadening applications to other facets of database management. Continuous evaluation and adaptability in diverse environments remain crucial, pushing the boundaries of how we perceive and optimize databases.

References

[1] Michael Felderer and Jeffrey C. Carver. *Guidelines for Systematic Mapping Studies in Security Engineering*. 2018. arXiv: 1801.06810 [cs.SE].

[2] "Guidelines for Conducting Systematic Mapping Studies in Software Engineering". In: *Inf. Softw. Technol.* 64.C (Aug. 2015), pp. 1–18. ISSN: 0950-5849. DOI: 10.1016/j.infsof.2015.03.007. URL: https://doi.org/10.1016/j.infsof.2015.03.007.

[3] Jonas Heitz and Kurt Stockinger. "Join Query Optimization with Deep Reinforcement Learning Algorithms". In: *CoRR* abs/1911.11689 (2019). arXiv: 1911.11689. URL: http://arxiv.org/abs/1911.11689.

[4] Zoi Kaoudi et al. "ML-based Cross-Platform Query Optimization". In: *36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020*. IEEE, 2020, pp. 1489–1500. DOI: 10.1109/ICDE48307.2020.00132. URL: https://doi.org/10.1109/ICDE48307.2020.00132.

[5] Barbara Kitchenham and Stuart Charters. "Guidelines for performing Systematic Literature Reviews in Software Engineering". In: 2 (Jan. 2007).

[6] Ryan Marcus et al. “Bao: Making Learned Query Optimization Practical”. In: *SIGMOD Rec.* 51.1 (2022), pp. 6–13. DOI: 10.1145/3542700.3542703. URL: https://doi.org/10.1145/3542700.3542703.

[7] Ryan Marcus et al. “Neo: A Learned Query Optimizer”. In: *Proc. VLDB Endow.* 12.11 (2019), pp. 1705–1718. DOI: 10.14778/3342263.3342644. URL: http://www.vldb.org/pvldb/vol12/p1705-marcus.pdf.

[8] Kai Petersen et al. “Systematic Mapping Studies in Software Engineering”. In: *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*. EASE’08. Italy: BCS Learning & Development Ltd., 2008, pp. 68–77.