SUOMEN AKATEMIA
FINLANDS AKADEMI • ACADEMY OF FINLAND

APPLICATION 325117

Submitted: 01.10.2018 11:28
Last supplemented: 06.11.2018 22:02

Applicant Michael Mathioudakis	Site of research and department University of Helsinki, 0313471-7 , Faculty of Science(Consortium MLDB)
Call Academy Project Funding 10.09.2018 - 01.10.2018 16:15	Research Council Research Council for Natural Sciences and Engineering

Research topic/Project title

Model Management Systems: Machine learning meets Database Systems (MLDB)

Model Management Systems: Machine learning meets Database Systems (MLDB)

Funding period 01.09.2019 - 31.08.2023	Funding to be applied for 792 766 €	Overall cost estimate 1 132 552 €

Keywords

database systems, data management, machine learning

database systems, data management, machine learning

Field of research

1. Computational data analysis
2. Computer science

Abstract

Research in the area of Database Management Systems (DBMSs) has focused for years on the efficient querying and processing of data. These efforts have led to the development of systems that implement general-purpose, highly optimized, and standardized technologies that allow users of DBMSs to focus on writing queries without having to grapple with the details of how (potentially large amounts of) data are organized physically on storage or how computational resources are managed by the system. The most prominent example of this concept is the Structured Query Language (SQL) and Relational DBMSs optimized for SQL queries.

Moreover, there has been massive interest in Machine Learning (ML) techniques that focus on the use of data for predictive tasks. However, unlike DBMSs, the landscape of ML systems remains fragmented and far from standardized. As a result, besides executing predictive tasks, users of ML systems are confronted with low-level decisions --- including the choice and processing of data, as well as the specification and training of statistical models for each predictive task, which are typically discarded after the completion of the task. This workflow wastes not only user effort but also computational resources. An alternative approach would be to build ML systems that treat statistical models as first-class citizens, to be maintained, optimized, and re-used by the system for indefinite time --- and separate system usage from management.

Our project aspires to develop a computational framework and associated techniques towards model management systems, i.e., systems that extend traditional DBMS functionality to allow users to focus on specifying predictive tasks with limited direct involvement in lower-level decisions for the management of data, models, and computational resources. The prototype system to be developed in this project is implemented as a software layer between DBMS and user interface. Internally, the system implements techniques for building model-based views and indexes for predictive queries, lazy incremental learning, and human-in-the-loop functionalities for data and model curation.

Research plan

1 Aim and objectives

1.1 Significance of the research project in relation to current knowledge

This project aims to develop a computational framework and associated techniques towards **model management systems**, i.e., systems that extend traditional DBMS functionality, to allow users to specify predictive tasks with limited direct involvement in lower-level decisions for the management of data, models, and computational resources. To accomplish this goal, the project builds upon and draws inspiration from research in both Database Management Systems (DBMSs) and Machine Learning (ML).

Research on DBMSs has focused for years on the efficient organization and querying of data. These efforts have led to the development of systems that implement general-purpose, highly optimized, and standardized technologies. They allow their users to focus on expressing their information requests in the form of structured queries, without requiring from users to grapple with details about how data are organized physically on storage, or how computational resources are managed by the system.

The most prominent achievement of such efforts is the development of the Structured Query Language (SQL) and Relational DBMSs (RDBMSs) optimized for SQL queries [RGE00]. In the SQL/RDBMS setting, users are responsible for formulating queries in a formal language (SQL), while the system manages physical data storage and organization, memory access, concurrency, and query execution, for which it offers guarantees (e.g., consistency for concurrent operations) and optimizations (e.g., selection of efficient execution plans). While the inner workings and organization of a DBMS remain hidden from the user, the DBMS makes certain design choices available to a system administrator --- who can, for example, request the materialization of indexing structures or instruct the system to favor certain execution plans over others. A similar approach is followed in non-relational (NoSQL) DBMSs (e.g., MongoDB) and distributed data-processing systems (e.g., Apache Hadoop or Spark) --- and has facilitated wide user adoption.

Research in machine learning has recently delivered major breakthroughs in predictive performance for a number of difficult tasks, including image recognition, speech recognition, natural language processing and translation, and gaming (e.g., software programs that beat humans in board games such as chess and go). Many of these breakthroughs are due to advances in the performance of deep neural networks [LBH15]. At the same time, increased computing power and advances in sampling and variational methods have renewed the field of Bayesian modeling and made probabilistic programming feasible in practical settings [CGH17]. As research in machine learning advances fast at many fronts, machine-learning software is becoming more widely adopted by the industry to deliver statistical insights from *big data,* in domains such as healthcare, medicine, finance, and commerce. The need to not only query data, but also to use data to extract insights has given rise to the term *data science* and demand for *data scientists* in the labor market, i.e., personnel with expertise in both data management and machine learning.

One area where machine-learning techniques find new adoption is that of DBMSs, as machine-learning approaches are explored to optimize or even replace components of DBMSs. For example, recent works suggest elaborate machine-learning approaches to develop indexes that are based on machine-learning models in place of traditional B-trees or hash-based indexes [KBC18, MIT18]; optimize join-query execution plans, traditionally produced based on ‘naive’ models that made independence assumptions for attributes of the data [CYI17, KHB17, KYG18]; or deliver continuously improving approximate query processing, traditionally based on static data synopses [PTC17].

At the moment, there is a large number and variety of **machine-learning systems** that are used in the industry. They include in-memory software libraries (e.g., general-purpose libraries such as Scikit-learn; or specialized ones, such as *Stan* for Bayesian analysis, *Tensorflow* for neural networks, or *Edward* and *Tensorflow Probability* for probabilistic programming). They also include machine-learning libraries that run on top of cluster-computing platforms (e.g., Apache Spark MLLib, Apache SystemML, Amazon SageMaker and Amazon Machine Learning, Google Bigquery ML and ML-engine, Microsoft Machine Learning Studio).

The aforementioned systems differ in many aspects, including the programming language and the settings they are tailored to. However, they can be seen as variants of a basic machine-learning pipeline, which we illustrate below with an example.

Example. Consider a data scientist who has access to the database of the products, customers, and purchases of an online store. The data scientist intends to use the data to build machine learning models that will allow her to perform *predictive queries*. Specifically, her ultimate goal is to get answers to predictive queries such as: “what is the probability that a customer in Helsinki would spend more than 200 euros to buy a desk”. To be able to answer such queries, she performs the following steps (see Figure 1): first, she submits a query to the DBMS that hosts the database so as to retrieve a subset of data that are relevant to her task; she then specifies a machine-learning algorithm to train a machine-learning model on the retrieved data; the output of the algorithm is an instance of the specified model; she finally uses this model to answer her predictive query.

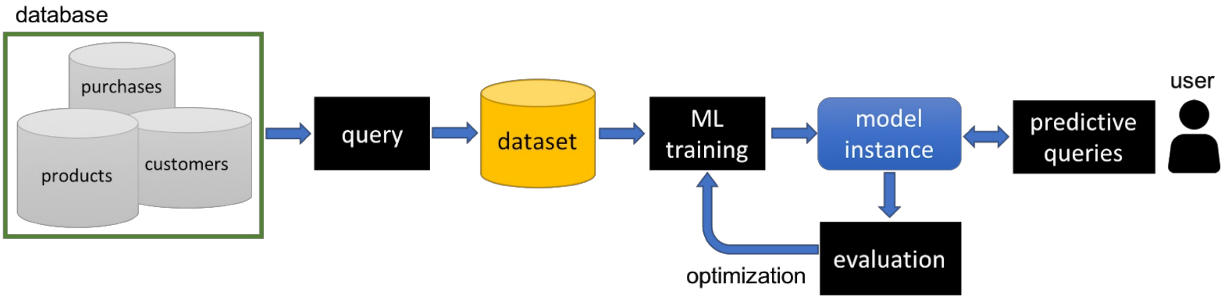


Figure 1: Typical ML pipeline. The user is heavily involved in all steps of the pipeline (stages represented with black boxes).

In the above scenario, which is all too typical in many systems, the system-user's expertise is required at all levels, from data management to model training to model usage: essentially the user has to be both a data-management and machine-learning expert. Comparing this setting with popular DBMSs, such as *Postgres* or *MySQL*, we observe that in the latter case users are only required to know how to formulate their queries.

At the same time, the system offers little help in optimizing the use of computational resources. For example, typically, machine-learning systems do not offer a good way to reuse the training effort towards one model instance, either by the same user or others who are interested in the same analysis --- and discard trained models by default, unless the user chooses to store them and reuse them. This is wasteful as oftentimes the user, whether a machine-learning expert or not, will repeat part of the pipeline to evaluate the model performance, e.g., the user may repeat the process for different datasets. This is particularly wasteful in settings where data grow: in such cases, re-training models from scratch, instead of incrementally, can add significant computational overhead.

Ideally, one would wish training efforts for the same or similar tasks to be re-used by the system in an automatic and optimal way.

Motivated by such observations, there have been recent efforts towards data and computational resource management for machine-learning systems, and prominent researchers in the DBMS community argue for the need of research in this direction [KMR16]. Some recent works have focused on specific directions for better management of computational resources. Specific examples of such directions include: materializing and re-using Machine Learning models [HTA18]; optimizing Linear Algebra operations for big-data settings [BRH18]; using sampling in a systematic way to allow system users to perform analysis at various levels of accuracy and speed [PQS18]; cost efficiency considerations for Gradient Descent [KQT17]; declarative query languages for machine learning [LCC17]; development of DBMS for large-scale array computations [BPG10], which are essential for today's neural-network-based algorithms; ML libraries on top of data processing systems (MLLib on Apache Spark) [MBY16]; predicting performance of ML tasks for cost optimization [VYF16]; and model versioning systems for deep learning models [MLD17a, MLDb].

All aforementioned work is very recent, aligns well with the focus of this project, and appears at first-tier, international venues. Even though this theme of research is very timely and impactful, and Finnish research institutions have a strong presence in Machine Learning research, there is little to no involvement of Finnish research in this research topic. We aspire to be the first Finnish consortium to work actively on this topic.

1.2 Theoretical premise

This project develops a computational framework and associated techniques towards machine-learning-powered database systems that allow users to perform predictive tasks with limited direct involvement in lower-level decisions for the management of computational resources. Such systems should be seen as extensions of existing DBMSs, that provide additional management for machine-learning models.

The MLDB project has the following objective: **develop a computational framework and associated methods to design and build systems that perform automated and optimized management of data and models under limited computational resources**.

The theoretical premise of our project is that such automated optimization is feasible under an appropriate definition of the involved resource management costs and the functionality which we aim to provide. We are encouraged in this belief by the success of the DBMS literature towards systems that offer automated optimization for execution plans of data queries.

Specifically, we envision systems that implement the pipeline shown in Figure 2. In this pipeline, *users* of the system are responsible only for formulating predictive queries they submit to the system. In addition, an *administrator* makes basic design choices for the system (e.g., what data are available for predictive tasks or what machine-learning models can be trained) and provides feedback for the curation of data and models via an AI-assisted mechanism offered by the system. Finally, the *system* is responsible for training ML models, materializing and maintaining ML models, optimizing execution plans for predictive queries for minimum use of computational resources, and offers a mechanism for the administrator to provide feedback. The system can be seen as extending the data management functionality of traditional DBMSs with model management.

The prototype system to be developed in this project is implemented as a software layer between the DBMS and the user interface (see Figure 2). Internally, the system implements techniques for three types of functionality, each addressed in its own work packages (WP).

WP1: Model-based views and indexes for predictive queries. Similarly to the view and indexing structures supported by traditional DBMSs, the proposed MLDB system implements structures that support frequently-occuring predictive tasks. Specifically, the system aims to perform automatic pre-computation (a.k.a. `offline' computation) that allow it to produce efficient answers to the values of joint and conditional probability distributions that are requested by the users at query time.

WP2: Lazy incremental learning: motivated by the huge volume of data in typical modern databases, training of machine learning models (i) is performed incrementally with data updates -- i.e., as data are inserted or deleted from the database; (ii) adapts lazily to queries that are submitted to the system -- i.e., the system aims to achieve good accuracy for the subspace of data that is relevant to the predicted queries that are submitted by the users; (iii) aims for reuse of previously trained models -- i.e., it aims to reuse previously trained models so as to minimize the amount of computational resources devoted to training at query time.

WP3: Human-assisted data-quality analysis. The system provides a human-in-the-loop functionality that enables a system administrator to guide model maintenance. Based on human feedback, the system adapts model accuracy to areas of interest and maintains a sequential lineage of models, each level corresponding to one round of human feedback.

An additional work package **WP4** focuses on the implementation of the aforementioned functionality into a prototype MLDB system.

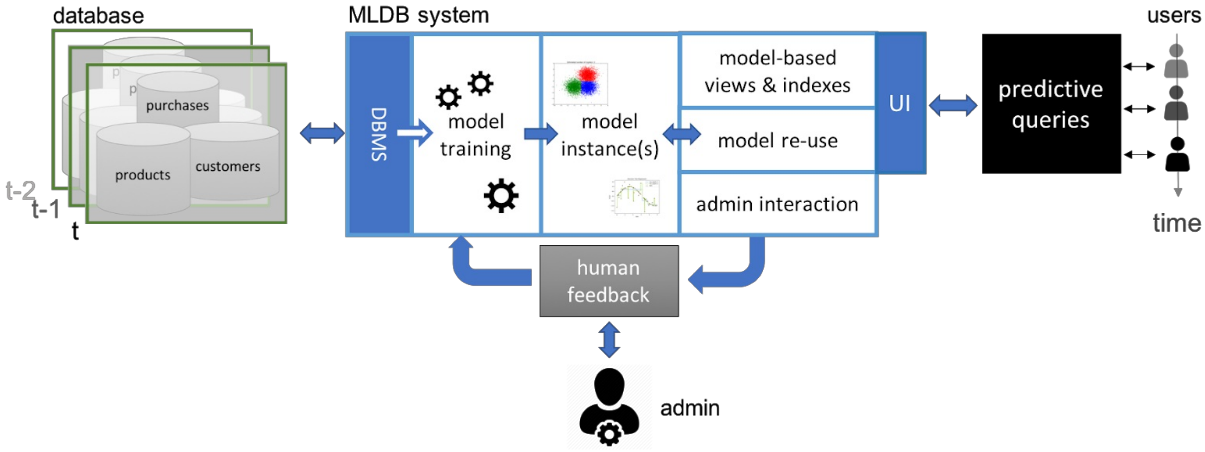


Figure 2: MLDB pipeline. In the suggested pipeline, users interact with the system only to submit predictive queries. An administrator makes basic design choices about the system and provides AI-assisted human feedback for better model training. The system can be seen as an extension to DBMS systems, that manages data as well as models, and adapts to dynamic changes in data and query workloads.

1.3 Hypotheses or research questions

The type of systems envisioned in this project will find application in settings that satisfy the following conditions: (i) one or more users use the system to perform predictive tasks on overlapping parts of the data using the same models; (ii) data change dynamically; (iii) the predictive tasks that are of interest to users vary with time. Thus, our project relies on the following hypotheses:

Hypothesis 1 Such settings do occur in practice.

Hypothesis 2 Intelligent management of resources in such settings would lead to significant computational savings.

Hypothesis 3 Automated and optimized management of resources is feasible.

Hypothesis 1 is supported by common knowledge and experience that the PIs have acquired from their own research efforts, research literature that addresses the challenges of one or more aspects of these settings, as well as discussions in the Database Management and Machine Learning community (for example, the public discussion by Prof. Joaquin Vanschoren et.al. at the DEEM Workshop at SIGMOD 2018, summarized at http://wp.sigmod.org/?p=2454).

Hypothesis 2 is supported by the results of recent research efforts, many of which we discussed in Section 1.1. For example, the work by Prof. Gautam Das et.al. [HTA18] reports significant speed-ups from the optimized re-use of materialized Machine Learning models.

Hypothesis 3 is also supported by the results of previous research efforts. For example, Prof. Matthias Boehm et.al. in [BRH18] demonstrate that it is feasible to produce fusion plans, i.e., execution plans that avoid redundant computation, for large Linear Algebra operations that are at the heart of many ML algorithms. Of course, even though previous work has produced evidence of this hypothesis, there is still a risk that the problem instances we will face are harder than the ones discussed in the literature. In particular, we should make sure that the algorithms we develop provide good-quality solutions in running times that make them usable in practice, and can be used to effectively simplify the tasks of the end users of the ML pipeline, as described earlier. We further discuss risks in Section 2.7.

Under the aforementiontioned hypotheses, we further specify the objective of the project in terms of the following Research Questions (RQs), each of them answered in a different work package.

RQ1 (WP1): What are the necessary pre-computations for an MLDB system to answer predictive queries fast for a given model instance trained from the data?

RQ2 (WP2): What are the necessary algorithms for an MLDB system to implement lazy incremental learning?

RQ3 (WP3): How can an MLDB system decide what human feedback is necessary to build better models over the data?

RQ4 (WP4): What are the main systems considerations for implementing an MLDB prototype and what is the methodology required for evaluating its performance and validating the overall concept?

1.4 Expected research results and their anticipated scientific impact, potential for scientific breakthroughs and for promoting scientific renewal

The novel contributions of this project are: (i) The definition and development of a new computational framework that is novel and timely in the context of both the machine learning and database-systems literature. (ii) The development of new methods to solve computational problems that have only recently attracted the focus of the research community. (iii) The development of a prototype MLDB system.

In terms of expected results, we plan to publish 3 journal papers, 12 regular conference papers, and 3 demonstration papers at major Data Management and Machine Learning venues. The papers correspond to the work packages and related research questions described above. They are listed in *Section 2.6 Schedule*.

With respect to technical quality of results, we will measure our impact in terms of improvement in performance of algorithms and systems. Performance is measured as the running time required to answer queries in a given workload, further specified according to the task. For tasks related to WP1/RQ1, we measure the time required for the system to answer predictive queries. For tasks related to WP2/RQ2, we measure the time for the system to produce a ML model that is adequeate for a given query workload. For tasks related to WP3/RQ3, we measure: (i) the number of interactions between computer and human until the system develops a model of desired accuracy, or (ii) the time and space required by the system to maintain and summarize the lineage of model versions built on human feedback.

With respect to wider scientific impact, we will measure success on the basis of (i) quality of doctoral dissertations, followed by success in the career paths of the PhD students; (ii) career path of the postdoctoral researchers; (iii) number and impact of publications in first-tier journals and conferences; and (iv) importance of the findings of the project; (v) release of software, which will be adopted by other scientists and practitioners. Target journals for disseminating the research developed in the project are IEEE TKDE, ACM TKDD, DMKD, etc. Target conferences are NIPS, ICML, VLDB, SIGMOD, KDD, WWW, etc.

2 Implementation

2.1 Data to be used

We'll be using two types of data. Firstly, publicly available **real data** available on online dataset repositories, including but not limited to once hosted by big computing companies such as Amazon (registry.opendata.aws/), Microsoft (docs.microsoft.com/en-us/azure/sql-database/sql-database-public-data-sets) and datasets discoverable through Google search (toolbox.google.com/datasetsearch). Secondly, **synthetic data** generated with the TPC benchmarking software (www.tpc.org/), widely used in the Data Management literature to compare the performance of algorithms. For more information, see the Data Management Plan (appendix).

The data will be used solely to evaluate the performance of the algorithms we develop, in terms of speed and quality. Given the focus of the project on computational performance, we will *not* be using the data to extract or publish whatever insights one might discover from the data. This definitely precludes the possibility that we publish sensitive data (e.g., personal or private) in the context of the project.

Data created by consortium members within the research project (e.g., synthetic data generated with the TPC software) and used for the evaluation of the developed algorithmic techniques, along with related software code and documentation, will be published by the time an analysis report is published based on the data. Software code that is essential for the processing of data and their analysis will be made publicly available on the Gitlab installations (gitlab.com) of the two universities, under MIT or GNU licenses.

2.2 Research methods

In this section, we discuss in more detail the work packages of the proposal and describe the relevant research methods and some of the directions we will pursue.

WP1. The goal of this work package is to automate the computation of data structures that facilitate the efficient answering of probabilistic queries. The kind of problems we will be solving can be seen as analogous to the problem of efficient materialization of data cubes presented by Harinarayan et al. [HRU96]. In that work, the authors describe how to construct data cubes, i.e., structures that contain aggregate values of a database-table attribute *y* for each combination of values of a set of other attributes X. To allow for efficient computation of data cubes for *arbitrary* attributes *X* and *y*, Harinarayan et al. explain how to pre-compute and materialize a small number of data cubes, from which other data cubes can be efficiently materialized at query time.

We are interested in performing a similar task, but in a probabilistic setting: compute the probability distribution of an attribute y for a given combination of values of a set of other attributes X. The pair (*X,y*) of attributes defines a *probabilistic query*. In the setting of the example introduced in Section 1.1, one such probabilistic query would ask for the probability that a customer would *buy a desk* given that the desk costs 200 euros and the customer is a resident from Helsinki. In probabilistic terms, we are looking for the probability

P(Customer-buys-desk = True | Cost-of-desk = 200, Location-of-customer = Helsinki).

In this case, the attributes of interest are *Customer-buys-desk* as *y*, and {*Cost-of-desk*, *Location-of-customer*} as *X*.

In practice, one computes this probability distribution by training a regression or clasification model for *y* conditional on *X*. However, training such a model for every arbitrary set of attributes *X* and *y* is extremely inefficient. A better approach would be to pre-compute a small number of such models that allow the efficient extraction of regression models for arbitrary attributes *X* and *y*. Such models can be seen as the probabilistic analogue of materialized view data structures that have been studied extensively in the Database Management literature. We are already working on developing algorithms to perform such pre-computations for Bayesian Networks, i.e., in settings where a Bayesian Network is used to capture the joint distribution of database attributes, but where a user submits probabilistic queries that involve an arbitrary subset of all attributes. We aim to submit this work to PVLDB 2019. In the future, we will extend it for settings of dynamically changing data and workload of queries. The challenge there will be to decide whether it is cost-efficient to update the pre-computated data structures, taking into account the trade-off between the computational cost to repeat the pre-computations, on one hand, and the loss of model accuracy if the system does not repeat them.

WP2. In the typical machine-learning pipeline, as shown in Figure 1, one or more users might build models on the same or overlapping parts of the data. This is wasteful: ideally, the system should be able to use previous model training to train new models more efficiently. To achieve such a model re-utilization objective, the system would have to choose among different training strategies: one strategy would be to train a new model from scratch, irrespective of any previously trained models; another strategy would be to combine the results from previously trained models with limited new training. The optimal strategy would be the one that leads to minimal use of computational resources (e.g., execution time). The problem is addressed in a recent PVLDB paper by Hasani et al. [HTA18], in which the authors describe an algorithm for optimal training strategy at query time, for a specific type of models (k-means models in one dimension). Significantly more research would be needed to extend the work of Hasani et al. to more general settings --- including settings of different models, dynamically varying data, and varying query workloads --- as well as to perform more efficient training strategy computation. One immediate step to extend the work of [HTA18] is to develop algorithms for multi-dimensional settings -- i.e., how to re-use k-means models that have been built on different regions of the (multi-dimensional) space of data.

WP3. In the ML pipeline that we envision (Figure 2), a system administrator (MLDB admin) will be providing concise feedback to the MLDB system --- not only to curate noisy data (e.g., correct errors or complete missing values), but also to make basic design choices, e.g., related to the level of accuracy that the various models or model-based views and indexes should attain. This latter task is in line with the task described by Park et al. [PQS18], in which an administrator sets the level of model accuracy. In the same paper, the authors describe a sampling-based system to adapt efficiently to different accuracy levels. We will extend such approaches to (i) allow the MLDB admin to set different accuracy levels to different parts of the data, (ii) develop an AI mechanism by which the system, in active-learning manner, requests miniman feedback from the admin, on order to set the levels of accuracy at different parts of the data. Every time the MLDB admin provides feedback, the models and model-related data structures are updated, leading to a lineage of models. We will also develop algorithms to maintain space- and time-efficient model lineages. This is in line with the task of "ML provenance" described by Kumar et al. [KMR16].

WP4. We will build a prototype system that implements the algorithms developed in WP1-3. The system will be implemented as a software layer on top of a relational DBMS, with three sublayers (one for the algorithms of each WP).

2.3 Human resources

The project consortium consists of two research groups in the University of Helsinki and Aalto University. The composition of the teams is as follows.

The University of Helsinki team will be led by **assistant professor Michael Mathioudakis, PI**, who will also be the project coordinator and responsible for the data management. The University of Helsinki team will recruit two researchers for this project, postdoctoral researcher P1 and doctoral student D1.

The Aalto University team will be led by **professor Aristides Gionis, coPI**. The team will employ two researchers, postdoctoral researcher P2 and doctoral student D2. Postdoctoral researcher P2 will be **Dr. Cigdem Aslay** who is already a member of the Aalto University team.

The PI and coPI will supervise the researchers and be involved in all WPs. Each one of the researchers will work mainly in one WP and will be responsible for that WP. In particular, Dr. Aslay (P2) will be responsible for WP1, D1 for WP2, D2 for WP3, and P1 for WP4. The doctoral students D1 and D2 will also contribute in the prototype work package WP4. Additionally, the team members will be involved on other WPs based on discussions, project meetings, skills, and interests, so as to leverage synergies and build a collaborative team environment.

2.4 Collaborators

Both the PI and coPI have extensive collaboration networks. Overall they have publications with more than 150 coauthors. The key players in the context of MLDB are the following:

Prof. Joaquin Vanschoren, Eindhoven University of Technology, works on automating machine learning and making it open and collaborative. He will collaborate with the consortium on WP1, and will host a research visit of Dr. Aslay (P2) for 3 months in 2020.

Prof. Gautam Das, University of Texas at Arlington, is a distinguished university chair professor and he has done fundamental research in all aspects of data management, including databases, data analytics and mining, information retrieval, and algorithms. In MLDB he will contribute in WP2, and will host a research visit of D1 for two months in 2020.

Dr. Gianmarco De Francisci Morales, research leader at ISI Turin, provides expertise on scalable data mining. He has collaborated extensively with the PI, the coPI, and P2 (Aslay). He will contribute in WP3 and he will host a research visit of D2 for two months in 2021. Additionally, he will host professor Gionis as part of an ISI fellowship (funding provided by ISI).

Prof. Matthias Boehm, Graz University of Technology, conducts research on data management systems and has strong expertise on building ML systems for declarative, large-scale machine learning. He will collaborate with the consortium on WP4. He will host a research visit of P1 for two months in 2021.

In addition to the research visits, we will organize a mini-workshop in Helsinki, in 2021, where we will invite all the external collaborators, in order to disseminate the results of the project and receive feedback. The mini-workshop will also enable to strengthen existing ties and identify directions for future work.

Table 1 summarizes the assignment of the MLDB team members and the external collaborators to work packages and the research visits.

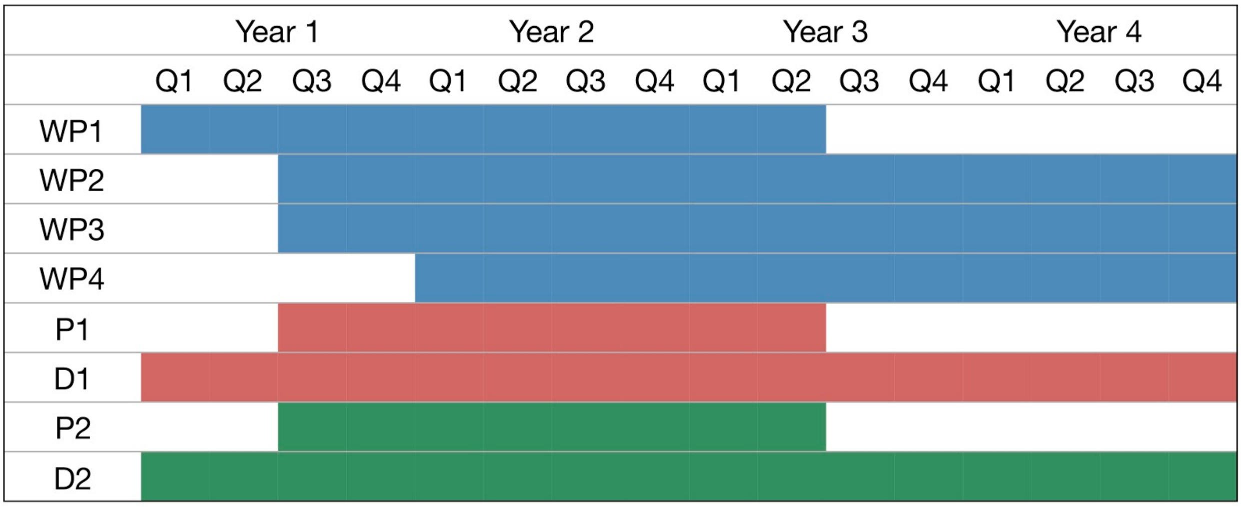
Table 1: Assignment of MLDB team and external collaborators to the work packages and planning of research visits.

Work package	MLDB assignment	External collaborator	Year of Research visit
WP1	Aslay (P2)	Prof. Joaquin Vanschoren, TU/e	2020
WP2	D1	Prof. Gautam Das, UoT, Arlington	2020
WP3	D2	Dr. De Francisci Morales, ISI Turin	2021
WP4	P1	Prof. Matthias Boehm, Graz U	2021

2.5 Research environment

The host institutes for MLDB are the Departments of Computer Science of the University of Helsinki and Aalto University. With a wide range of computing resources and support services, a truly international community, and commitment to high-quality research and teaching, both Universities provide excellent supporting environments. Both universities occupy a strong position in European and global research arenas in computer science, as demonstrated by their high placement in international university rankings. The departments of Computer Science in both Universities have long-standing tradition in the areas of machine learning and data management, and they attract a continuous stream of bright PhD students and postdoctoral researchers from Finland and abroad.

2.6 Schedule



	Year 1				Year 2				Year 3				Year 4			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
WP1	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦						
WP2			🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
WP3			🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
WP4					🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
P1			🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥						
D1	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
P2			🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩						
D2	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

Figure 3. WP1: Model-based views and indexes; WP2: Model update and re-use; WP3: Human in the loop; WP4: MLDB System Prototype; P1/D1: University of Helsinki researchers; P2/D2: Aalto University researchers.

The project is planned to run for 4 years from September 2019 to August 2023. We divide the work into four packages as discussed previously. The work will be split roughly equally among the two research teams. Work packages WP2 and WP4 are led by the University of Helsinki team, while work packages WP1 and WP3 are led by the Aalto team.

The four work packages will run partially concurrently, and they will span almost the whole duration of the project. This is essential as we plan to advance the work by iteratively applying the results obtained from one task to re-evaluate and improve the methods of the other dependent tasks. The time schedule of the project is shown in Figure 3.

The two teams will be in close collaboration. Monthly meetings will be organized, centered around the study of related work and presentations of preliminary results obtained within the project. In addition to conference traveling, all researchers will visit the project collaborators, for a time period of two or two months. We expect four such mobility trips, as shown in Table 1.

With respect to publications, our plan is shown in Table 2. In addition to the topic we also mention the target venue and time for the first submission.

Table 2: Publication plan

WP	Topic	Target	Year
WP1	Model-based views and indexes for Bayesian Networks and other joint-probability models (*in progress*)	PVLDB	2019
WP1	Updates of model-based views and indexes in the presence of dynamic data	SIGMOD	2020
WP1	Updates of model-based views and indexes in the presence of dynamic workloads	PVLDB	2021
WP1	Efficient maintenance of model-based views and indexes	TKDE	2022
WP2	Missing data and missing models: a Bayesian approach for the re-use of approximate models	NIPS	2020
WP2	Model-reuse in the presence of dynamic data	ICML	2021
WP2	Model-reuse in the presence of dynamic workload data	NIPS	2022

WP2	Efficient model re-use with guarantees for the construction of approximate models	JMLR	2023
WP3	A human-in-the-loop approach for active learning of target accuracy levels of ML models	NIPS	2021
WP3	Model Version Control: A framework for the maintenance of machine-learning model versions in the presence of human feedback	PVLDB	2022
WP3	Choosing what to forget: efficient summarization of model version lineage	SIGMOD	2023
WP4	MLDB: a system for efficient predictive queries	SIGMOD demo	2021
WP4	MLDB: an optimizer for model re-use	SIGMOD demo	2022
WP4	MLDB: a system that learns target accuracy levels via human interaction	SIGMOD demo	2023

2.7 Risk assessment and alternative implementation strategies

MLDB sets ambitious goals and has high potential, but at the same time it contains risks. Below we identify the highest risks of the project, we quantify their likelihood and their potential impact, and we suggest the measures we will take to mitigate those risks.

Risk 1. Project hypotheses: A potential risk is that we will find out that some of the project hypotheses cannot be validated, namely it is not beneficial to design and build systems that perform automated and optimized management of data and models under limited computational resources. There may be different reason for this, such as that relevant scenarios do not occur in practice, or that the existing machine-learning pipeline can handle the majority of real-world applications without the need to leverage model-based views or indexes, neither lazy incremental learning, nor human feedback in order to build better models over the data. Following the arguments presented in our motivating discussion, we believe that this risk is low. Furthermore, even if the project concept cannot be universally validated, we believe that there will be several important real-world application scenarios that our framework will provide significant benefits. Additionally, we believe that the project concept provides the ground for developing novel ideas from the academic-point-of-view.

Risk 2. Algorithms: Devising efficient algorithms with provable quality guarantees is a challenging task, which, however, gives the largest potential for scientific impact in the computer-science community. In cases that we will not be able to prove theoretical results, we will study problems with simplifying assumptions, and we will focus on devising heuristic methods and providing thorough empirical validation.

Risk 3. MLDB prototype: We aim to build a prototype system that implements the algorithms developed in the project. The system will be implemented as a software layer on top of a relational DBMS. Building such a system requires significant effort, and recruiting expertise with strong software development skills. We will indeed take into account this consideration during the recruitment process. However, due to the scope and the size of the project this risk has high likelihood. The mitigating action will be to release the software that we will develop as a non-integrated suite of methods and seek support from the open-source software community.

Risk 4. Datasets: A common risk in many data-driven projects is the availability of data. In this project, however, our concept is not tied to any specific dataset and we can guarantee success by validating our methods on synthetic and benchmark datasets. Thus, in MLDB the data risk is non-existent.

Overall, the project is structured in a way that there are medium-risk paths, which will still make it highly successful. Furthermore, succeeding on the high-risk tasks will make the project groundbreaking.

2.8 Added value of consortium

The added value of a consortium in DBML is substantial; in fact the project could not be adequately executed without bringing together a complementary set of skills and exploiting the synergies between the two groups. Prof. Gionis' group contributes their expertise in algorithms design, theoretical aspects of machine learning, and discrete optimization. The group has worked on a number of different projects on developing data analysis and algorithmic techniques to database applications and has published several papers in top database venues such as PVLDB. Prof. Mathioudakis' group, on the other hand, provides expertise on probabilistic modeling aspects as well as in database systems. The group has also significant expertise in developing prototypes of applications that require large-scale data processing and analysis. The two groups have collaborated extensively in the past, and ongoing collaboration includes work along the themes of this project. A joint consortium allows for efficient investigation into designing model management systems, formulating problem abstractions, developing efficient solutions, and building prototypes to evaluate the proposed solutions in practice. By necessity, the project will run in an iterative fashion and in close collaboration between the two groups during each development cycle, which will be further facilitated by the close geographical proximity of the two hosting institutions.

3 Responsible science

3.1 Research ethics

Research will be carried out in compliance with the European Code of Conduct for Research Integrity. Any data used in the project will be reviewed for compliance with the EU general data protection regulation (GDPR). Furthermore, as we will use only synthetic data and datasets that have already been processed for publication, we do not expect any ethical issues regarding data usage.

3.2 Promotion of open science

All the data and materials that will be used in DBML will be publicly available. Our publications will be available via open access, in particular they will be shared via the arxiv.org repository. Similarly, the software we will develop and the other outputs of the project will become freely available to the scientific community via the institutional GitLab installations (gitlab.com) of the University of Helsinki and Aalto University.

3.3 Promotion of equality and non-discrimination

During the hiring process we will support diversity and consider actions to avoid discrimination and achieve gender balance. Indicatively, in the current team of prof. Gionis the gender ratio is 3:5 while each person comes from a different country. Prof. Mathioudakis joined the University of Helsinki a few months before the proposal was submitted and has not made a recruitment yet, however, equality and non-discrimination considerations will be taken into account.

4 Competence of research team and collaborators

4.1 Merits of research team members that are relevant to the project

The research team consists of the applicant, Assistant Professor Michael Mathioudakis, as the PI, Professor Aristides Gionis as the coPI, two postdoctoral researchers (P1, P2) and two doctoral researchers (D1, D2). Researcher P2 is Dr. Cigdem Aslay, who is currently a postdoctoral researcher in Aalto. Researchers P1, D1, and D2 will be recruited via open calls. During recruitment we will aim to cover the needs of the research project, bringing expertise both in machine learning and database systems research.

PI **Michael Mathioudakis** is an assistant professor at the University of Helsinki, and Chair for Algorithmic Data Science at the Helsinki Center for Digital Humanities (HELDIG). Prior to that, he was a Postdoctoral Researcher at Aalto University and received his PhD from the University of Toronto. His past research has focused on the algorithmic analysis of user generated content on the Web, with a recent emphasis on online polarization and algorithmic fairness. He also has experience in teaching Database Management Systems and Machine learning courses at postgraduate level --- and is currently teaching Data Management and AI applications at Aalto's Professional Development program (http://bit.ly/aaltoproAI).

coPI **Aristides Gionis** is a professor in the department of Computer Science in Aalto University. His previous appointments include being a visiting professor in the University of Rome and a senior research scientist in Yahoo! Research. His contributions span several areas of data science, including algorithmic data analysis, web mining, social media analysis, data clustering, and privacy-preserving data mining. He is currently serving as an action editor of the *Data Mining and Knowledge Discovery* journal, an associate editor the *ACM Transactions on Knowledge Discovery from Data*, and an associate editor of the *ACM Transactions on the Web*. He has served in the programme committee of numerous premium conferences, including being the PC chair for WSDM 2013 and ECML PKDD 2010. His work has received several best-paper awards at international conferences. His papers have over 15000 citations, and his *h*-index is 52 (Google Scholar).

P2 **Cigdem Aslay** (PhD) is a postdoctoral researcher at Aalto University. Her work focuses on algorithmic methods for graph mining and social network analysis, with an emphasis on social influence propagation in online social networks. Her PhD dissertation constitutes one of the first algorithmic investigations in the intersection of social influence propagation, viral marketing and social advertising.

The team is in a unique position to accomplish the goals of the project. The PIs bring together theoretical work on machine learning and database systems, which provide the foundations for this project. One of the unique strengths of the team, as demonstrated by past research, is to combine successfully theory and practice, by formulating novel problems, designing scalable algorithms with provable quality guarantees, and applying these methods in relevant applications. The practical aspects of the project are further enhanced by the six-year experience in industrial research by one of the PIs (Gionis in Yahoo! Research). The PIs have an extensive network of international collaborators, four of which will be collaborating directly in this project.

4.2 Merits of collaborators that are relevant to the project

The external collaborators of the project are prof. Gautam Das from the University of Texas at Arlington, prof. Joaquin Vanschoren from the Technical University of Eindhoven, prof. Matthias Boehm from the Graz University of Technology, and Dr. Gianmarco De Francisci Morales from ISI Foundation, Turin. They all have excellent track record and they all push the state-of-the-art in their respective areas, thus, we expect them to provide strong support on the project tasks and objectives. We have invited them as external collaborators due to existing ties with the team and/or the relevance of their work with MLDB. All four external collaborators are working on areas that are in the intersection of machine learning and database management. We have also selected them for their complementary skills and profiles. Prof. Das is more senior and well accomplished, while the other three are in earlier steps of their career. Prof. Das and prof. Boehm bring more focus on database research, while prof. Vanschoren has more expertise in machine learning research. Dr. De Francisci Morales has a very good balance between data mining and database systems research. Furthermore, the team has established collaborations with prof. Das and Dr. De Francisci Morales, while prof. Boehm and prof. Vanschoren will be new collaborators. We expect from our external collaborators to host members of the project in their home institutions and co-author papers. We also plan to invite them all for a project meeting in Helsinki in the mid-term of the project, where we will present them our preliminary results and obtain feedback for the project continuation.

4.3 How the project is linked to previous research by the PI or the research team, or to some other research?

The team has not done previous research on problems related to model management systems and incorporating machine-learning capabilities in database-management systems, and in this respect the project opens up a new area of interest for the PIs. However, the PIs have pioneered research in many areas in machine learning and databases that have strong connections with the project theme from the methodological point of view. Such work includes problems in probabilistic modeling, handling uncertainty in data, model selection, clustering problems, and optimization. The PIs are regularly publishing in the major journals and conferences in machine learning, data mining, and database management, such as, journals TKDE, DMKD, TKDD, and conferences PVLDB, KDD, ICDM, ICDE, and more.

5 Societal effects and impact

5.1 Effects and impact beyond academia

In addition to scientific publications and software distribution, more opportunities will be sought to disseminate the output of the project to a wider audience and improve societal impact. From an educational point-of-view, both PIs are often invited for tutorials in conferences, which can provide a forum to disseminate the results of the project within the academic community and the industry. Additionally, we will seek opportunities to make our work accessible to a wider audience; notably, our recent work has received coverage from media outlets or popular blogs (research.cs.aalto.fi/dmg/media.shtml). Furthermore, we will explore opportunities to collaborate with companies and start-ups on tasks related to the project themes; both hosting institutions provide an excellent environment to incubate such collaborations. Finally, the project will increase its societal impact through doctoral training and by advancing the career of young researchers. The research group of prof. Mathioudakis is newly formed, but has secured funding for 1 PhD student and 1 research assistant --- while prof. Gionis has already graduated 5 doctoral students and 4 postdocs who are currently employed in high-tech companies and academic institutions, such as, Google, Microsoft, Nokia Bell Labs, ISI Foundation, EPFL, and more.

5.2 Considering principles of sustainable development

Our project is in the broad area of information and communications technology aiming to automate machine-learning tasks and provide model-management capabilities in database systems. As such, the project will provide innovative methods to manage data more effectively and more efficiently. The methods of the project can be used in a wide variety of applications, for example, applications related to knowledge and information access, participatory society of citizens, wellbeing, and improvement of local communities. It follows that the project can provide support for a sustainable society of information and knowledge. Furthermore, as our methods directly aim to improve efficiency of computation and optimal allocation of available resources, the project will help to save energy and reduce carbon footprint.

6 Bibliography

6.1 Provide a list of all the sources you have used in the research plan

[BPG10] Brown PG. Overview of SciDB: large scale array storage, processing and analysis. InProceedings of the 2010 ACM SIGMOD International Conference on Management of data 2010 Jun 6 (pp. 963-968). ACM.

[BRH18] Boehm M, Reinwald B, Hutchison D, Sen P, Evfimievski AV, Pansare N. On optimizing operator fusion plans for large-scale machine learning in systemML. Proceedings of the VLDB Endowment. 2018 Aug 1;11(12):1755-68.

[CGH17] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. Stan: A probabilistic programming language. Journal of statistical software. 2017 Jan 1;76(1).

[CYI17] Chen Y, Yi K. Two-level sampling for join size estimation. In Proceedings of the 2017 ACM International Conference on Management of Data 2017 May 9 (pp. 759-774). ACM.

[GTK01] Getoor L, Taskar B, Koller D. Selectivity estimation using probabilistic models. In ACM SIGMOD Record 2001 May 1 (Vol. 30, No. 2, pp. 461-472). ACM.

[HTA18] Hasani S, Thirumuruganathan S, Asudeh A, Koudas N, Das G. Efficient construction of approximate ad-hoc ML models through materialization and reuse. Proceedings of the VLDB Endowment. 2018 Jul 1;11(11):1468-81.

[HRU96] Harinarayan V, Rajaraman A, Ullman JD. Implementing data cubes efficiently. In Acm Sigmod Record 1996 Jun 1 (Vol. 25, No. 2, pp. 205-216). ACM.

[KBC18] Kraska T, Beutel A, Chi EH, Dean J, Polyzotis N. The case for learned index structures. In Proceedings of the 2018 International Conference on Management of Data 2018 May 27 (pp. 489-504). ACM.

[KHB17] Kiefer M, Heimel M, Breß S, Markl V. Estimating join selectivities using bandwidth-optimized kernel density models. Proceedings of the VLDB Endowment. 2017 Sep 1;10(13):2085-96.

[KMR16] Kumar A, McCann R, Naughton J, Patel JM. Model selection management systems: The next frontier of advanced analytics. ACM SIGMOD Record. 2016 May 9;44(4):17-22.

[KQT17] Kaoudi Z, Quiané-Ruiz JA, Thirumuruganathan S, Chawla S, Agrawal D. A Cost-based Optimizer for Gradient Descent Optimization. In Proceedings of the 2017 ACM International Conference on Management of Data 2017 May 9 (pp. 977-992). ACM.

[KYG18] Krishnan S, Yang Z, Goldberg K, Hellerstein J, Stoica I. Learning to optimize join queries with deep reinforcement learning. arXiv preprint arXiv:1808.03196. 2018 Aug 9.

[LBH15] LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015 May;521(7553):436.

[LCC17] Li X, Cui B, Chen Y, Wu W, Zhang C. Mlog: Towards declarative in-database machine learning. Proceedings of the VLDB Endowment. 2017 Aug 1;10(12):1933-6.

[MBY16] Meng X, Bradley J, Yavuz B, Sparks E, Venkataraman S, Liu D, Freeman J, Tsai DB, Amde M, Owen S, Xin D. Mllib: Machine learning in apache spark. The Journal of Machine Learning Research. 2016 Jan 1;17(1):1235-41.

[MIT18] Mitzenmacher M. A Model for Learned Bloom Filters and Related Structures. arXiv preprint arXiv:1802.00884. 2018 Feb 3.

[MLD17a] Miao H, Li A, Davis LS, Deshpande A. Towards unified data and lifecycle management for deep learning. In2017 IEEE 33rd International Conference on Data Engineering (ICDE) 2017 Apr 1 (pp. 571-582). IEEE.

[MLD17b] Miao H, Li A, Davis LS, Deshpande A. ModelHub: Deep Learning Lifecycle Management. InData Engineering (ICDE), 2017 IEEE 33rd International Conference on 2017 Apr 19 (pp. 1393-1394). IEEE.

[PQS18] Park Y, Qing J, Shen X, Mozafari B. BlinkML: Approximate Machine Learning with Probabilistic Guarantees.

[PTC17] Park Y, Tajik AS, Cafarella M, Mozafari B. Database learning: Toward a database that becomes smarter every time. InProceedings of the 2017 ACM International Conference on Management of Data 2017 May 9 (pp. 587-602). ACM.

[RGE00] Ramakrishnan R, Gehrke J. Database management systems. McGraw Hill; 2000.

[VYF16] Venkataraman S, Yang Z, Franklin MJ, Recht B, Stoica I. Ernest: Efficient Performance Prediction for Large-Scale Advanced Analytics. In NSDI 2016 Mar 16 (pp. 363-378).

7 Other information entered in the research plan

7.1 How researcher training will be organised and research careers promoted in the project

The project will strive to develop the academic experience of doctoral students and postdoctoral researchers hired by the project.

Doctoral students hired by the project will work towards publications listed in Section 1.4: “Expected results and impact”, under the guidance of the group leaders (i.e., the PI and coPI of the project), as well as the postdoctoral researchers of the project. They will also take the role of Research Assistant for courses and seminars organized by the two group leaders. Doctoral researchers hired by the project will be co-instructors for those courses, which will be part of the Master’s programs offered by the University of Helsinki and Aalto University..

Both doctoral students and postdoctoral researchers will receive financial support to attend international conferences to present the project's work, and they will participate in research mobility (they will visit the external collaborators of the project). They will also be assigned as advisors for Master's theses at the two universities.

7.2 PI's working hours and salary in the project

The project is not requesting a salary for the PI. The PI's salary is covered by the salary he receives at the University of Helsinki. The time that the is required to allocate on research activities is more than enough to cover his role in this project.