Protocol for a Systematic literature review on distributed data management in Machine learning systems

AWDAY KORDE(VRIJE UNIVERSITEIT AMSTERDAM / UNIVERSITEIT VAN AMSTERDAM)

Protocol for a Systematic literature review on distributed data management in Machine learning systems

ABSTRACT

With the emerging field of Machine Learning, many data management frameworks have been emerged that address stream or batch data processing. These frameworks provide continuous processing, aggregation and analysis of bounded or unbounded data. Typically, the type of data processing is context bounded, in machine learning this is concerned with the type of problem the model addresses. While some models are trained across a finite collection of historical data, some require real-time processing where the values provided would be ingested in a continuous flow by receivers and separated into mini-batches. This paper addresses the issues encountered when implementing these systems into real use cases and the difficulties they come across when implementing data-intensive applications. Specifically, the papers that discuss the entire end-to-end work-flow around batch-processing machine learning models, the application design when constant aggregation of data and complex queries are required.

KEYWORDS

Distributed data management systems, batch-processing, stream-processing, machine learning systems, bounded data, un-bounded data.

Contents

List of Figures

List of Tables

1 Background and rationale

Distributed data management solutions have become a must when dealing with a continuous stream of information flowing in enterprise-grade systems. Migrating a relational monolith database to a microservice-type database has arrived with its own benefits and challenges. Martin Fowler defines that almost all successful microservice story has started with a monolith that got too big and broke up.

His remarks emphasize on the complexity of starting off with a microservice design at first, specifically, when envisioning the end system as a divided block of services. This common misinterpretation usually comes along with its own challenges which can be mitigated by iteratively ensuring that each aspect of the system has its own context bounded objective.

Within the emerging field of machine learning, data-intensive systems are built around pretrained models that provide them the necessary scheduled re-training, data governance, loggings and so forth. Currently, the architectural design of a system that would maintain these analytical models are diverse. Each model having it's own method of processing the incoming data, either be batch or stream processing. The methods in which a machine learning systems is designed highly depends on the data source the analytical model deals with, while some leverage online data (scheduled crawlers that constantly update or appends new data into underlying databases) or data coming from a web application that provides information on the products that have been ordered or purchased. In this Systematic Mapping Search (SMS) we aim at identifying the existing research on the limitation imposed when architecting machine learning systems. This goal is of finding, classifying and evaluating the current limitations of machine learning systems from different perspectives, such as industry adopted approaches, research-oriented strategy and so forth.

1.1 Existing systematic studies on the topic

In 2015, ODonovan et al. [1] performed a systematic mapping study on big data technologies used in manufacturing. The study looked at industry-wide and research-wide practices of big data technologies, specifically on identifying the areas in which it has been applied. By aggregating a number of 661 publications, ODonovan et al. study analyzed the frequency to contribution types in which the publications he collected have contributed to the development of big data technologies over time. The first filtering of the aforementioned publications was achieved by removing the articles or journals that did not clearly show their contribution to the big data field in manufacturing. Our study, somewhat similar, differs from the one done by O'Donovan et al. [1] as follows: (i) big data technologies have indeed play an important role when addressing distributed data management systems, in our study the primary focus is on how these technologies can be used together to allow seamless integration between them. (ii) O'Donoval et al. study touches upon a broad industry-wide application of these technologies, our focus is that of diving into their applicability with machine learning models. (iii) Our systematic search query and topics touch upon the techniques in which systems communicate with each other, more specifically, how synchronous or asynchronous communication imply some form of limitation to the system design.

Sinoara et al. study[2] in 2017, focused on the text semantics and natural language topics in publications where text mining was the method of extraction. The main points of this study was that of: (i) identifying and classifying of 3984 semantic-concerned publications into seven different usage aspects (ii) provide a thorough overview of the field such that future researchers and/or expert would take advantage of. Our study differs from the one Sinoara et al. performed, their main focus was that of finding how latent knowledge extraction methods are currently used in the semantic web as well as text mining in combination with the prior.

Both aforementioned studies provide a broad representation of the technologies involved or the process of retrieving textual information respectively. Unlike this study, the priors studies don't address the possible limitations of system due to the increased implementation complexity of such systems.

1.2 The need for an SMS on current implementation of machine learning systems

With this systematic mapping study we aim at unveiling some of the most common limitations when architecting machine learning systems. This will help researchers and practitioners identify and design the main setbacks their end-system might encounter. Furthermore, we will look into the current research that has been done with the aim of improving the current techniques of architecting machine learning systems with a keen focus on consistency, availability and usability. By doing so, experts within the field of machine learning will have a sense of direction regarding the end-state of their system. Industry-wide experts can make use of this study to identify and classify each system that is concerned with at least one machine-learning functionality and provide a point of reference when designing their end-system.

2 Research Process

In this paper the systematic search will be carried out by using the process shown in Figure 1; it is divided into three main phases reviews [3, 4]: planning, conducting, and documenting.

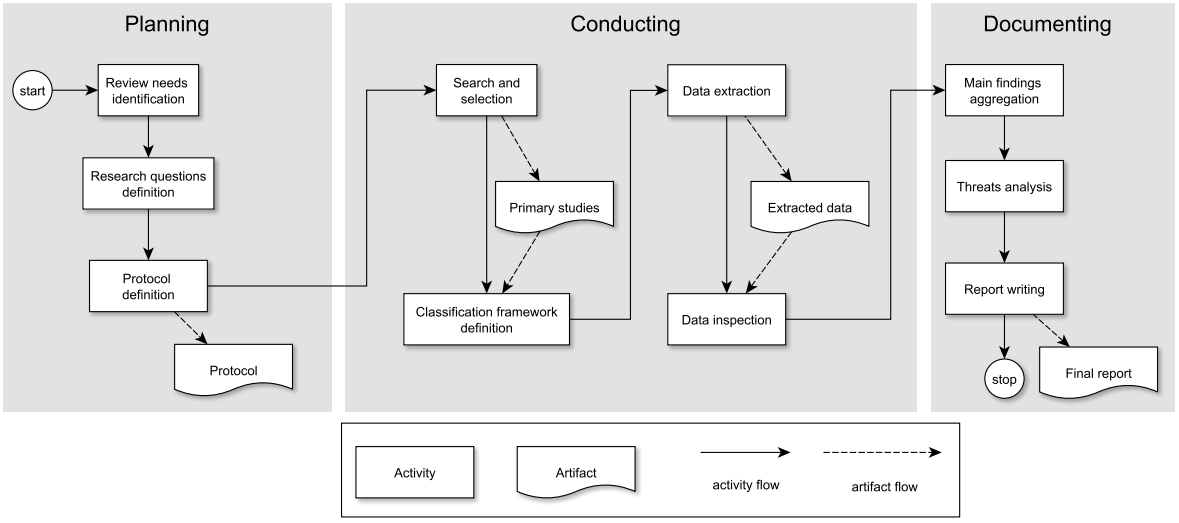


Figure 1: Overview of the whole review process

These phases represent the stages of research during this study, each having a specific purpose that will help us identify and filter the publication in an iterative manner.

2.0.1 Planning

This phase aims to (i) establish the necesassity of having a study that approaches the topic proposed (see Section 1.2), (ii) identify the research questions this study aims to answer (see Section 3).

2.0.2 Conducting

In this phase we will perform the mapping study by following all the steps previously defined in the review protocol. More specifically, we will carry out the following activities:

- *Search and selection*: In this activity, by using a set of filters and inclusion and exclusion criteria, we will flag potential publication that address data management in machine learning systems.In section 4 we will elaborate the selection process as well as the techniques involved in doing so.

- *Comparison framework definition*::In this step, we will identify potential candidates that specifically address the limitations imposed by machine learning systems. These candidates will define the set of features that are going to be used as features for further steps.
- *Data extraction*: in this activity we will dive into the studies identified and fill in the corresponding data extraction form. Afterwards, the forms will be collected, analyzed and aggregated such that we can further analyze it in the next step.More details about this activity are presented in Section 5.
- *Results*: This activity will focus on addressing the research questions by making use of the aggregated and analyzed data from the previous step to arrive at a meaningful conclusion. The details about this activity are presented in Section 7.

2.0.3 Documenting

Within the section of the paper we will address the following activities: (i) a detailed overview of the data extracted from the prior phase which going to be elaborated in a result-centric manner. (ii) A summary of the report describing the result along with each activity that contributed to the findings.

3 Research questions

The purpose of this study is that of finding state of the art systems for validating the purpose of this study, that is, the limitation imposed by implementing such systems. The two main target groups the result of this study is going to focus on are as follows: (i) Researchers that are about to or are currently implementing a machine learning system of their own and want to do so by knowing the issues encountered so far by others. (ii) Experts and practitioners that deal with such systems on a daily basis or are willing to adopt new techniques of architecting machine learning systems.

In the following section we are going to elaborate on the specific goal by using a Goal-Question-Metric perspectives(i.e., purpose, issue, object, viewpoint) . Where for each of the aforementioned perspectives are going play a role in identifying and classifying the characteristics of each publication extracted. The abstract goal of this study is going to be phrased into research questions along with it's relevance to this study as follows:

RQ1: *What are the **publication trends** of research studies about architecting machine learning systems?*

Rationale: In academic and industry-wide research, nowadays, there are numerous researchers and experts within the field of machine learning that end up building entire systems that would support their model. With this questions we aim at unveiling the frequency of publications that address machine learning and distributed databases. Based on the trends over the past years, we aim at determining the amount of occurring publications that touch upon the aforementioned topic.

RQ2: *What is the **focus** of research in machine learning when architecting highly-scalable systems?(e.g. a publication that might propose a highly-scalable online-learning systems, where factors such as, consistency or synchronous updates of a global data set)*

Rationale: Scalable machine learning systems usually are separated into either batch-processing or stream-processing. Looking at the aspects they are focused on, we can identify the diversity of topics touched upon within each of these methods of processing data. By doing so, we can properly classify and identify the current research gaps on architecting scalable machine learning systems.

RQ3: *What are the **potential limitations or flaws** of machine learning systems in distributed environments?*

Rationale: With this research question we aim at extracting the rationale behind each publication which can be easily transferred to industry-wide practices. Ultimately, we are specifically interested in the work-around or solutions that state of the art systems provide. With this study we look into the experiments done and the components that led to functionality impediments when distributed databases was addressed.

The research questions will drive the whole study, with a special influence on (i) search and selection of primary studies, (ii) data extraction, and (iii) data analysis.

4 Search and selection process

In this section, we will elaborate on our search processes and selection protocols. The main idea behind this study is of retrieving publications that carry some sort of relevance towards our field of interest, specifically machine learning.

More specifically, it is fundamental to achieve a good trade-off between the coverage of existing research on the topic considered, and to have a manageable number of studies that are to be analyzed. In order to achieve the aforementioned mentioned trade-off, our search strategy consists of an automatic search. This allows us to have control over the filter and query process, while still being able to include multiple inclusion and exclusion criteria's. The automatic search process is as follows:

1. Initial search. The aim of this stage is that of performing an automatic search on an electronic database(Table 1). For the purpose of this study, we choose only one database, IEE Xplore Digital Library which is part of the most largest and most complete scientific database in software engineering [5].

Name	**Type**	**URL**
IEEE Xplore Digital Library	Electronic database	http://ieeexplore.ieee.org

Table 1: Electronic database considered in this research

For the purpose of this study, the search string focuses specifically on our topic of interest. Furthermore, by making use of the RQ we defined in the previous section we will further narrow down the search along the way.

```
("machine learning" AND "distributed" AND (stream* OR batch*) AND
syst* AND (sync* OR async*))
```

Listing 1: Search string used for automatic research studies

2. Impurity removal. In the second phase of the search process, the results retrieved by the query mentioned in the prior section were promising. Some of the publications retrieved were books, workshop proceedings, articles and so forth. Our focus here is that of retrieving research papers that were relevant to our study, so we can draw a meaningful conclusion based on them. We removed all the papers that were not research papers.

3. Merging and duplicates removal. In this phase, duplicate threads will be merged, this is done by matching their title,author,year and venue of publication.

4. Application of selection criteria. After we successfully removed the impurities and merged all the duplicate publications, the process of filtering the papers begins. We define a set of well-defined criteria which we are going to make use as steps of the filtering process. These criterias

are broken down into two categories: Inclusion and Exclusion criterias. These categories are well-defined such that we have an objective base-line towards which paper might qualify for this study. The outcome of these criterias will enable us to label each publication as relevant or irrelevant to the purpose of this study. The full-text reading of each paper will follow after the inclusion and exclusion has been applied to all the publications retrieved.

5. Combination. In this section, we will keep track of all the papers that have similar studies, where the combined outcome of similar papers point to the same outcome. This is ensures completeness and traceability of each resulting paper. (e.g. A primary publication that approaches the same topic of asynchronous communications between machine learning services which serves the same end-purpose as a different study. These publications will be bundled together such that we can keep track of the subjects addressed.)

4.0.1 Selection criteria

As mentioned by Barbara Kitchenham et al. [3] the process of defining the selection criteria is always when the protocol definition is established. This is done such that we can reduce the probability of bias when selecting primary studies that satisfy the inclusion criteria.

Inclusion criteria

I1) Studies that focus on machine learning systems(Reinforcement learning systems, Recommendation systems, Clustering systems etc..), specifically on the process of how these systems were built.

I2) Studies that discuss the technical implementation and type of architectures that are used in machine learning systems (e.g. batch-processing system that makes use of different components that are tightly-couple, where communications between components are synchronous.)

I3) Studies written in English.

Exclusion criteria

E1) Studies that, while focusing on machine learning, do no specifically touch upon how these components interact or communicate with each other. These are studies that specifically focus on the algorithmic aspect of machine-learning.

E2) Studies that touch upon machine learning-systems yet don't give any details regarding the technical implementation of such system.

E3) Studies that provide an general approach to distributed systems. More specifically, studies that address a generic aspect of implementing distributed applications which are not concerned with machine learning.

E4) Studies not available as full-text.

4.1 Classification of research

All primary studies were classified using three-dimensions. Each given dimension provides different perspectives on the current progress or direction of each specific study. The dimensions aforementioned are as followed:

1. Type of research. A scheme by Wieringa et al. [6] was proposed for classifying engineering research, seen in Table 2. This will enable this to identify the type of research that is being carried out by the authors.

Classification	Description
Validation	Publication that investigate novel and unique techniques
Evaluation	Publication that include implementation in practice of techniques. These studies usually evaluate the techniques of prior researchers and focus on the implementation process along with the outcomes and an evaluation
Solution	Publications that propose a solution to an existing problem. Solutions can be either an illustration, novel or an improved version of an existing solution.
Philosophical	Publication provides a different way of approaching a particular problem, usually in the form of structuring a problem into a new way
Opinion	Publication that provides a personal opinion regarding an certain technique or implementation, where the opinion does not rely on scientific proof or research methods
Experience	Publication that derives from the biography or experience of an author. Shows how the research has been done in practice, from the author's perspective

Table 2: Type of research strategies

2. Area of practice in machine learning. In order to identify the type of work in machine learning each publication is addressing, an existing classification schema is chosen, seen in Table 3. specifically addressing, in which field of practice do the authors apply machine learning. The scheme, defined by Meziane et al[7] represents different areas of industry in which Artificial Intelligence is applied.

Classification	Description
Design	Initial solutions developed during the design phase of products
Process and Planning	Solutions that on process planning phase of the system
Quality Management	Statistical process control and quality control for manufacturing management
Maintenance and Diagnosis	Pattern recognition in optimization systems
Control	In charge of feasibility analysis, reasoning, self-improvement and adaptation
Scheduling	Optimization of computational processes and task scheduling

Table 3: Areas where Artificial Intelligence has been applied

3. Type of contribution. By looking at the type of contribution of each publication we can determine the contribution of the research. We take a qualitative research method, *keywording*, where we highlight keywords that describe the contribution of the publication. In table 4 we show a summary for each type of scientific contribution.

5 Data extraction

The main aim of this activity is that of being explicitly into what kind of data extraction is necessary for answering each research question established(see Section 3).

For the purpose of this study, we addressed each question in separate sections. This is done to maintain a well-structured data extraction approach, such that each section will be addressed

Classification	Description
Architecture	Theoretical view that describes the logical layers of the system and their interaction
Framework	Research that focuses on the combination of multiple software-libraries aims to solve an existing problem or extend a already-existing one
Theory	Research that focuses on theoretical aspect of the problem, provides guidelines and sufficient generic information, while not considering applied-research
Methodology	Similar to the theory classification, but provides in-depth information about the approach proposed
Model	Research that focuses on an statistical analytical approach to solve problems
Platform	Research that focuses on the hardware-software interaction, mainly referring to the enabling aspect of applications
Process	Research that focuses on the process of solving a specific problem, provides a detailed description of each process involved
Tool	Research that aims at developing tools and frameworks that addresses specific problems

Table 4: Types of contributions

by each of the three different classification schemes. Each of these schemes will make use of the classification methods mentioned in Section 4.1. The structure of the data extraction and analysis is as followed:

1. *Publication trends* addressing RQ1: In this section we will adhere to the classification scheme defined in section 4.1, table 2,specifically we will collect data about the type of research carried by each publication. Firstly, We will determine the distribution of the total and primary studies that address machine learning systems and provide an overview of the type of journals they were published in.

2. *Focus of research* addressing RQ2: In this section we will further classify the primary studies according to the type of Machine learning area addressed in each publication(As described in section 4.1,table 3) as well as, the type of contribution towards research based on table 4.

3. *Potential limitations* addressing RQ3: In this section we will collect all the results extracted from RQ2 and pick the publications where the research strategy would be that of proposing a solution and type of contribution an architecture or framework. This is done such that we can identify publication that aim at improving the performance of machine learning systems in a distributed environment.

All of aforementioned activities will be documented and collected rigorously for each publication extracted. Each of the retrieved studies and extracted information will be stored in a spreadsheet record, such that we can represent and derive the necessary information accordingly.

6 Results

6.1 Publications trends (RQ1)

Before selection process As observed in Figure 2, a year-on-year increase in all publications that address machine learning in distributed systems. With the first publication identified from 1971, the amount of publications that addresses machine learning in distributed system has seen a slow increase from 1971 till the end of 2004. From 2005 onward, the amount of publications

has nearly doubled after each year, resulting an increase of nearly 150% between 2011 and 2012. Where the amount of publications has increased tenfold between 2011 and 2017, this is due to the gain in usage and applicability of machine learning among researchers. The sudden decrease between 2017 and 2018 is due to the insufficiency of data containing only publications collected from January 2018.

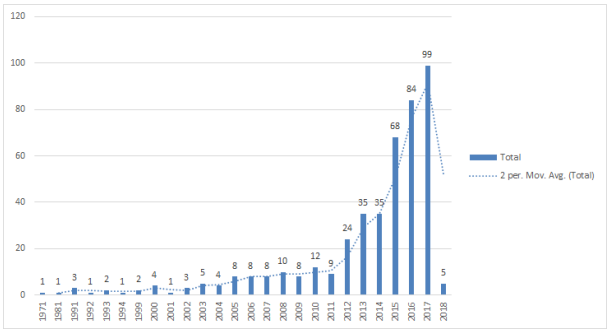


Figure 2: Total study distribution

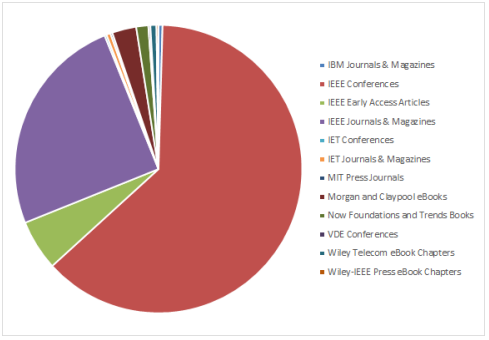


Figure 3: Overview of total studies journals

In Figure 3 we have a overview of all the studies by the type of journals and conferences they were published in. Among all, 63% of all the studies are published in IEEE Conferences, due to the nature of this study, we only extracted publications from IEEE Xplore Digital Library. Moreover, 25% of the studies have been published in Journals & Magazines. Where 6% of the total studies are IEEE Early Access Articles. All the content that is published under IEEE Early Access are articles that are made available before the final electronic versions are issued.

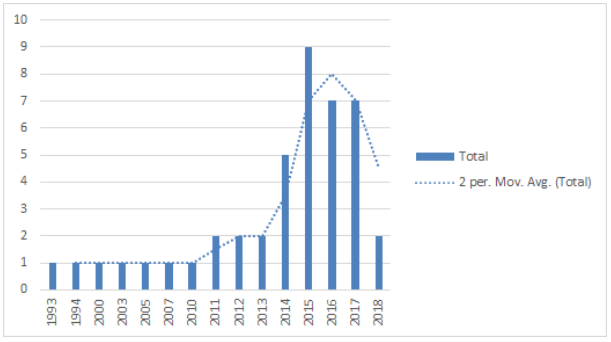


Figure 4: Distribution of studies

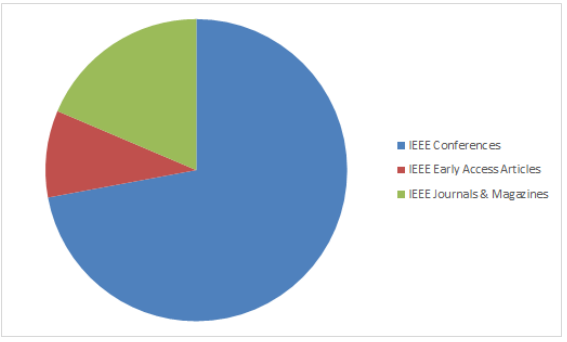


Figure 5: Total studies by journal

After selection process As seen in figure 4 the amount of studies that remain after we perform the inclusion and exclusion criteria reduced from 442 to 43 publications. These are the primary studies that address architecting machine learning systems, that being said, the proportion of articles and journals have increased to 300% between 2013 and 2015. An interesting observation is that between 1993 and 2010 there has been no increase of studies that address architecting machine learning systems, this is due to the focus of research on building efficient large-scale machine learning systems. As soon as interest in the area of machine learning started increasing, many publications started addressing the matter of how machine learning systems should be designed. This is due to the necessity of increasing the performance of large scale machine learning systems in production, such that the developed models would provide up-to-date predictions and could withstand the usage of multiple-users at the same time. Figure 5 highlights the distribution of primary studies by the journals they have been published in. 72% of the primary studies were published in IEEE Conferences, while 19% of the studies are published in IEEE Journals & Magazines.

6.2 Focus of research (RQ2)

After retrieving the primary studies the next step was of classifying the publication even further by using the classification schemes defined in Section 5. We first classified the publications by type of research, field of practice and type of contribution. As illustrated in Figure 7,the majority of the publication extracted are solution-based (65% of total primary studies) research, which is ideal for the purpose of this paper. Solution-based research aims at solving a specific problem as well as giving examples on how the proposed solution performs in practice. The next most prominent is the evaluation-based research (23% of total primary studies), where the authors propose an implementation and evaluates its performance by measuring the performance,accuracy or efficiency of the proposed-solution. Evaluation-based solutions are thoroughly explored by the author such that the all the drawbacks, benefits are illustrated for the proposed solution.The two least common types are Validation-based and Opinion-based research which compromise 9% and 3%, respectively, of the primary studies. As illustrated in Figure 7, the vast majority of publications address the area

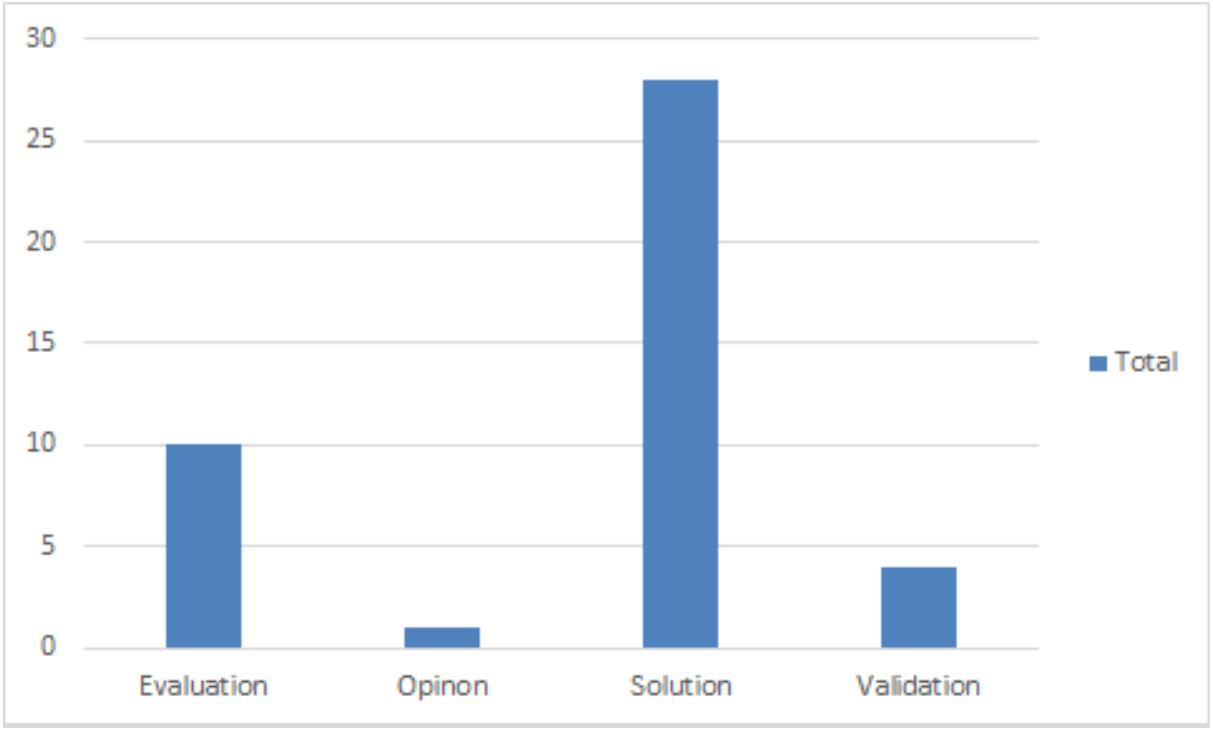
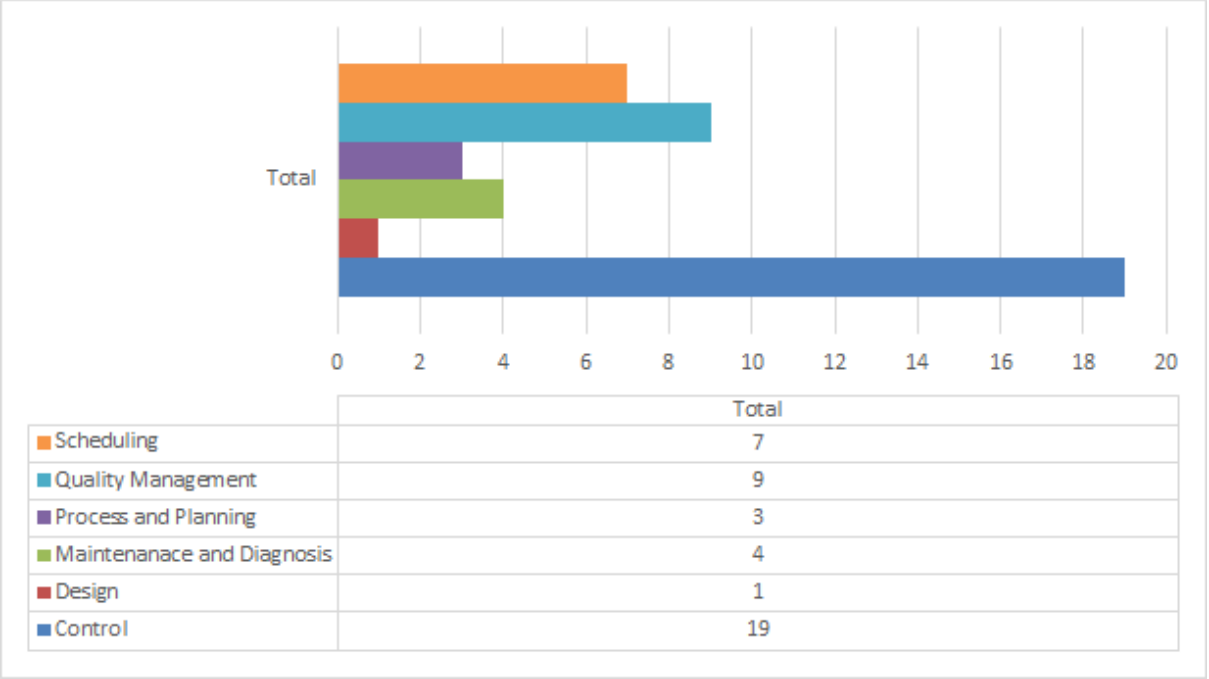


Figure 6: Primary study type

of control(44% of primary studies) in research where machine learning systems are concerned. The papers address topics such as online learning, reasoning engines, statistical inference and so forth. The second most prominent one is quality management, where publications address the correctness of data, data consistency, data availability and partition tolerance. The amount of publications that are focused on control are twice as many than than the publications that address quality management. This is due to the fact that there has been a sudden increase in systems that can support analytical inference, statistical reasoning ,image recognition owning to the computational power, cloud services, infrastructure speed that is more accessible and more powerful nowadays than it was before. The next most eminent field of practice in machine learning is that of process and planning, which comprises of 16% of total primary studies. Machine learning and statistical inference is used during the planning phase in the form of simulation and forecasting of unexpected events such that failure can be mitigated as much as possible. The applicability of machine learning in planning and process may also come in the form of building statistical models with the purpose of forecasting the amount of information that flows within a system such that the amount of running

processes can scale up or down based on the incoming workload. Figure 8 shows the distribution of



	Total
Scheduling	7
Quality Management	9
Process and Planning	3
Maintenanace and Diagnosis	4
Design	1
Control	19

Figure 7: Machine learning use case for each study

contributions made by each publication. The most prominent contribution are frameworks, where the research focus is that of creating end-to-end systems that are able to provide a seamless encapsulation between statistical models, data extractors, preprocessing pipelines and user interfaces. Most of the focus in this type of contribution is due to the necessity of having a reliable modular system that would be consist of multiple layer of processes. While frameworks comprise 35% of the primary studies, publications that provides statistical models (28% of primary studies) as contribution only approach the theoretical algorithmic aspect of machine learning. Where the purpose of the models would be either to improve the current implementation of the previous one or present a novel approach of a new algorithm for solving a particular problem. The third most prominent contribution is the theoretical type, most arguably the one that offers an in-depth high-level explanation of machine learning systems that are deployed in distributed environments.

6.3 Potential limitations (RQ3)

After successfully classifying each publication and retrieving the necessary information, the next step was that of diving into publications that addressed the subject of improving different components of machine learning systems functioning at a large-scale. A number of 18 publications were selected and further inspected to pinpoint the main flaws or limitation the authors were addressing. After thoroughly analyzing each selected publication, we can observe, based on the specific problem each publication was addressing, that the main problem the papers were addressing was that of avoiding a coordination behavior of components whenever possible such that an adaptive and flexible solution can be derived. While others address the process of model training that performs synchronous and cyclic operations.Moreover, these operations in a distributed environment can cause communication delays and ultimately slow down worker nodes that would otherwise perform better when a more optimized scheduling is in place. Another issue addressed in the papers is the replication strategies performed nowadays, specifically, database snapshots, materialized views and methods of optimizing partially replicated databases.

This is due to current strategies being 'fixed' and are unable to adapt to the types of data encoun-

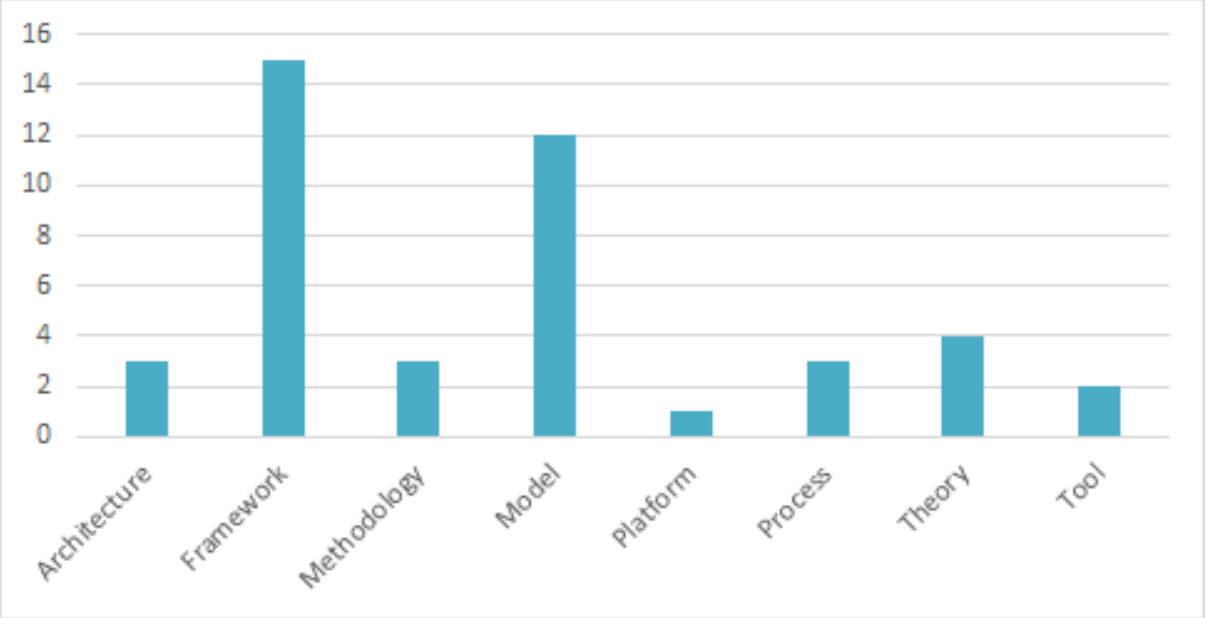


Figure 8: Primary study contribution

tered, thus requiring frequent expensive update of materialized views such that data consistency can be maintained. With the increasing necessity of have a real-time analytics system, the focus is that of providing up-to-date data consistency and availability. According to the CAP theorem [8], a distributed system can simultaneously able to provide two out of three options. The three options being consistency, availability and partition tolerance. The increasing popularity of machine learning in practice has shown that the current industry is moving towards providing systems that should be able to provide a high availability and adopt an eventual consistency model, where by adopting an incremental replication technique of database servers means that they will eventually be synchronized.

7 Threats to validity

With this systematic mapping study, we aim at mitigating potential risks that might affect the quality and reliability of the study. There are a number of threats that concern the validity of this study. In this section we will discuss the identified validity threats.

Search criteria
The search criteria that was utilized to retrieve the papers that addressed the topic of our study, was individually chosen. The search string was decided by (a) the topic we chose addresses specifics of machine learning in distributed systems specifically the publications that contained key-words related to specific data processing types and communication type. (b) There were multiple search strings, but based on multiple tests and the time-frame dedicated to this study, we decided to pick the search string that was the closest related to our study purpose. While the search string was carefully inspected before moving forward with this study, there might be many papers that address machine learning or distributed systems in a different manner. Examples such as, deep learning or online-learning instead of machine learning were omitted throughout the duration of this study.

Digital databases
The main digital database used to retrieve papers for this study was IEEE Xplore Digital Library. While this databases is part of the seven most well-known digital databases, there are multiple

digital databases that have been omitted such as, ACM Digital Library, Google Scholar, Scopus and henceforth. The aforementioned databases have been omitted due to time-constraints allocated for this study. This poses a validity threat to the content retrieved from publications, because there is a risk that relevant publications, which might have contributed to the end-results of this study, have not been added to this study.

Classification schema accuracy
In this study, we made use of three different classification schema to label each primary study. This was done such that we can identify the type of research, contribution and area of practice of each study. Considering that this study was done individually, the annotating and labeling was done by a single individual, which may differ from the perspective of a researcher. Due to this we might have introduced an individual bias toward the classification process of primary studies, with no subject to a review by another individual/researcher. This poses a threat of validity of accurately labeling the publications.

8 Conclusion and future work

Throughout the duration of this study, we results of this paper showed that there are multiple papers that address the limitations and flaws of machine learning systems in a distributed environment. At the time this study was conducted, this study was the only one that focused on systematically mapping the impediments encountered by researchers addressing specific problems in machine learning systems. Furthermore, several research questions that we addressed, specifically the distribution, trend and area of practice of publications throughout time, will pave the road map to further investigation regarding these occurring limitations.

9 Appendix

Table 5: List of primary studies

Title	**Year**	**Study Type**	**Area of practice**	**Type of contribution**
Distributed frank-wolfe under pipelined stale synchronous parallelism	2015	Solution	Scheduling	Framework
ScaAnalyzer: a tool to identify memory scalability bottlenecks in parallel programs	2015	Solution	Maintenanace and Diagnosis	Tool
Power system data management and analysis using synchrophasor data	2014	Evaluation	Quality Management	Methodology
Realizing Large-Scale Interactive Network Simulation via Model Splitting	2012	Solution	Maintenanace and Diagnosis	Model

TiDIntroducing and Benchmarking an Event-Delivery System for BrainComputer Interfaces	2017	Solution	Maintenanace and Diagnosis	Tool
Blazes: Coordination analysis for distributed programs	2014	Solution	Scheduling	Framework
Adaptive Bitrate Selection: A Survey	2017	Evaluation	Design	Theory
Infinite Factorial Finite State Machine for Blind Multiuser Channel Estimation	2018	Validation	Control	Model
Orchestrating safe streaming computations with precise control	2014	Solution	Quality Management	Framework
Ensemble Coordination for Discrete Event Control	2011	Validation	Control	Methodology
A 57 mW 12.5 J/Epoch Embedded Mixed-Mode Neuro-Fuzzy Processor for Mobile Real-Time Object Recognition	2013	Solution	Control	Platform
Self-Calibration: Enabling Self-Management in Autonomous Systems by Preserving Model Fidelity	2012	Evaluation	Control	Model
MatrixMap: Programming Abstraction and Implementation of Matrix Computation for Big Data Applications	2015	Solution	Quality Management	Framework
Industrial Analytics Pipelines	2015	Solution	Process and Planning	Architecture
Tessellation: Refactoring the OS around explicit resource containers with continuous adaptation	2013	Evaluation	Maintenanace and Diagnosis	Process
MemepiC: Towards a Unified In-Memory Big Data Management System	2018	Solution	Process and Planning	Framework

NOVADIB: a novel architecture for asynchronous, distributed, real-time banking modeled on loosely coupled parallel processors	1993	Solution	Control	Architecture
On the impact of virtualization on the I/O performance of analytic workloads	2016	Evaluation	Quality Management	Process
Learning Image and User Features for Recommendation in Social Networks	2015	Solution	Control	Model
GPSA: A Graph Processing System with Actors	2015	Solution	Control	Process
Stampede: a cluster programming middleware for interactive stream-oriented applications	2003	Solution	Control	Framework
Multimodal Affective User Interface Using Wireless Devices for Emotion Identification	2005	Opinon	Control	Theory
Panopticon: A lock broker architecture for scalable transactions in the datacenter	2015	Solution	Quality Management	Architecture
GPU in-Memory Processing Using Spark for Iterative Computation	2017	Solution	Scheduling	Framework
Adaptive Distributed Database Replication Through Colonies of Pogo Ants	2007	Solution	Quality Management	Framework
Spark Versus Flink: Understanding Performance in Big Data Analytics Frameworks	2016	Evaluation	Scheduling	Methodology
DPS: A DSM-based Parameter Server for Machine Learning	2017	Solution	Quality Management	Framework
High Performance Parallel Stochastic Gradient Descent in Shared Memory	2016	Solution	Control	Model

Efficient Distributed Machine Learning with Trigger Driven Parallel Training	2016	Solution	Scheduling	Framework
Leveraging MapReduce and Synchrophasors for Real-Time Anomaly Detection in the Smart Grid	2017	Validation	Control	Model
Temporal Distributed Learning with Heterogeneous Data Using Gaussian Mixtures	2011	Solution	Control	Model
A Scalable Multicore Architecture With Heterogeneous Memory Structures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs)	2017	Solution	Control	Framework
An Asynchronous Mini-Batch Algorithm for Regularized Stochastic Optimization	2016	Solution	Control	Framework
Automotive big data: Applications, workloads and infrastructures	2015	Evaluation	Process and Planning	Theory
A multi-layer software architecture framework for adaptive real-time analytics	2016	Solution	Quality Management	Framework
Column-Oriented Data Store for Astrophysical Data	2014	Solution	Quality Management	Theory
Continuously Improving the Resource Utilization of Iterative Parallel Dataflows	2016	Validation	Scheduling	Model
Scheduling the allocation of data fragments in a distributed database environment: a machine learning approach	1994	Solution	Scheduling	Framework
Mining tweets of Moroccan users using the framework Hadoop, NLP, K-means and basemap	2017	Solution	Control	Model

Non-strict cache coherence: exploiting data-race tolerance in emerging applications	2000	Evaluation	Control	Model
Asynchronous Adaptation and Learning Over NetworksPart III: Comparison Analysis	2015	Evaluation	Control	Model
Automatic extraction of topics on big data streams through scalable advanced analysis	2014	Solution	Control	Framework
Evidence updating for stream-processing in big-data: Robust conditioning in soft and hard data fusion environments	2017	Evaluation	Control	Model

9.0.1 Primary papers selected for RQ3

1. N. L. Tran, T. Peel S. Skhiri, Distributed frank-wolfe under pipelined stale synchronous parallelism,2015
2. P. Alvaro; N. Conway; J. M. Hellerstein; D. Maier,Blazes: Coordination analysis for distributed programs,2014
3. P. Li; K. Agrawal; J. Buhler; R. D. Chamberlain Orchestrating safe streaming computations with precise control 2014
4. Y. Huangfu; J. Cao; H. Lu; G. Liang MatrixMap: Programming Abstraction and Implementation of Matrix Computation for Big Data Applications 2015
5. K. E. Harper; J. Zheng; S. A. Jacobs; A. Dagnino; A. Jansen; T. Goldschmidt; A. Marinakis Industrial Analytics Pipelines 2015
6. Q. Cai; H. Zhang; W. Guo; G. Chen; B. C. Ooi; K. L. Tan; W. F. Wong MemepiC: Towards a Unified In-Memory Big Data Management System 2018
7. S. Ghosh NOVADIB: a novel architecture for asynchronous, distributed, real-time banking modeled on loosely coupled parallel processors 1993
8. U. Ramachandran; R. S. Nikhil; J. M. Rehg; Y. Angelov; A. Paul; S. Adhikari; K. M. Mackenzie; N. Harel; K. Knobe Stampede: a cluster programming middleware for interactive stream-oriented applications 2003
9. S. Tasci; M. Demirbas Panopticon: A lock broker architecture for scalable transactions in the datacenter 2015
10. S. Hong; W. Choi; W. K. Jeong GPU in-Memory Processing Using Spark for Iterative Computation 2017
11. S. Abdul-Wahid; R. Andonie; J. Lemley; J. Schwing; J. Widger Adaptive Distributed Database Replication Through Colonies of Pogo Ants 2007

12. C. Sun; Y. Zhang; W. Yu; R. Zhang; M. Z. A. Bhuiyan; J. Li DPS: A DSM-based Parameter Server for Machine Learning 2017

13. S. Li; J. Xue; Z. Yang; Y. Dai Efficient Distributed Machine Learning with Trigger Driven Parallel Training 2016

14. S. Moradi; N. Qiao; F. Stefanini; G. Indiveri A Scalable Multicore Architecture With Heterogeneous Memory Structures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs) 2017

15. H. R. Feyzmahdavian; A. Aytekin; M. Johansson An Asynchronous Mini-Batch Algorithm for Regularized Stochastic Optimization 2016

16. A. Vakali; P. Korosoglou; P. Daoglou A multi-layer software architecture framework for adaptive real-time analytics 2016

17. A. R. Chaturvedi; A. K. Choubey; Jinsheng Roan Scheduling the allocation of data fragments in a distributed database environment: a machine learning approach 1994

18. W. Romsaiyud Automatic extraction of topics on big data streams through scalable advanced analysis 2014

References

[1] P. O'Donovan, K. Leahy, K. Bruton, D. T. J. O'Sullivan, Big data in manufacturing: a systematic mapping study, Journal of Big Data 2 (1) (2015) 20. doi:10.1186/s40537-015-0028-x.
URL https://doi.org/10.1186/s40537-015-0028-x

[2] R. A. Sinoara, J. Antunes, S. O. Rezende, Text mining and semantics: a systematic mapping study, Journal of the Brazilian Computer Society 23 (1) (2017) 9. doi:10.1186/s13173-017-0058-7.
URL https://doi.org/10.1186/s13173-017-0058-7

[3] B. Kitchenham, P. Brereton, A systematic review of systematic review process research in software engineering, Information and software technology 55 (12) (2013) 2049–2075.

[4] C. Wohlin, P. Runeson, M. Höst, M. Ohlsson, B. Regnell, A. Wesslén, Experimentation in Software Engineering, Computer Science, Springer, 2012.

[5] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, M. Khalil, Lessons from applying the systematic literature review process within the software engineering domain, Journal of Systems and Software 80 (4) (2007) 571–583.

[6] R. Wieringa, N. A. M. Maiden, N. Mead, C. Rolland, Requirements engineering paper classification and evaluation criteria: A proposal and a discussion 11 (2006) 102–107.

[7] F. Meziane, S. Vadera, K. Kobbacy, N. Proudlove, Intelligent systems in manufacturing: current developments and future prospects, Integrated Manufacturing Systems 11 (4) (2000) 218–238. doi:10.1108/09576060010326221.
URL https://doi.org/10.1108/09576060010326221

[8] E. Brewer, A certain freedom: thoughts on the cap theorem, in: Proceedings of the 29th ACM SIGACT-SIGOPS symposium on Principles of distributed computing, ACM, 2010, pp. 335–335.