DOGO4ML: Development, Operation and Data Governance for ML-based Software Systems

Claudia Ayala, Besim Bilalli, Cristina Gómez and Silverio Martínez-Fernández[1]

[1] Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain

Abstract

Machine Learning based Software Systems (MLSS) are becoming increasingly pervasive in today's society and can be found in virtually every domain. Building MLSS is challenging due to their interdisciplinary nature. MLSS engineering encompasses multiple disciplines, of which Data Engineering and Software Engineering appear as most relevant. The DOGO4ML project aims at reconciling these two disciplines for providing a holistic end-to-end framework to develop, operate and govern MLSS and their data. It proposes to combine and intertwine two software cycles: the DataOps and the DevOps lifecycles. The DataOps lifecycle manages the complexity of dealing with the big data needed by ML models, while the DevOps lifecycle is in charge of building the system that embeds these models. In this paper, we present the main vision and goals of the project as well as its expected contributions and outcomes. Although the project is in its initial stage, the progress of the research undertaken so far is detailed.

Keywords

DevOps, Machine Learning, DataOps, Data and Software Engineering, ML-based software systems

1. Introduction

The European Political Strategy Centre stated that *"Data is rapidly becoming the lifeblood of the global economy. It represents a key new type of economic asset. Those that know how to use it have a decisive competitive advantage in this interconnected world, through raising performance, offering more user-centric products and services, fostering innovation—often leaving decades-old competitors behind."*[1]. It becomes necessary for companies to master the development, operation, and governance of software systems that embed advanced statistical models exploiting data for different purposes. Such models are typically generated using **Machine Learning (ML)**, i.e., *"the study of computer algorithms that improve automatically through experience"*, which rely on available sample data to learn models and *"make predictions or decisions without being explicitly programmed to do so"* [1]. We call **ML-based software systems (MLSS)** those software systems whose behavior is greatly determined by ML models embedded therein. MLSS are becoming increasingly pervasive in today's society and are present in virtually every domain: from smart mobility (autonomous driving) and Industry 4.0 (factory robots) to smart health (diagnostic systems) and smart infrastructures (cloud-based services), etc.

Processes for building MLSS tend to be complex, inherently iterative and difficult to manage and govern. One of the reasons for this complexity is that they encompass multiple disciplines, of which **Data Engineering (DE)** and **Software Engineering (SE)** appear as most relevant. In setting up MLSS, data and software engineers are often faced with several challenges that make even more complicated their development and operation: *(i)* the lack of a well-established set of good practices to design, manage and govern, in a systematic manner, such software systems; *(ii)* the increasingly usual

[1] https://ec.europa.eu/epsc/sites/epsc/files/epsc_strategic_note_issue30_trategic_autonomy.pdf

characteristics of Big Data, and *(iii)* the lack of definition of specific indicators and quality requirements for MLSS (e.g., related to trustworthiness or ethics), and tool support for validating them.

This paper presents the **DOGO4ML** project, acronym for Development, Operation and Data Governance for MLSS. DOGO4ML (https://dogo4ml.upc.edu/en) is a 4-year project that started on September 2021 and it is funded by the Spanish research agency, under the National Spanish Program for Research Aimed at the Challenges of Society 2020 (RETOS 2020).

DOGO4ML is run by the integrated Software, Services, Information and Data Engineering research group (**inSSIDE**, https://insside.upc.edu/) at the Universitat Politècnica de Catalunya (**UPC**). inSSIDE is composed of two subgroups: *(i)* the Software and Service Engineering group (**GESSI**, https://gessi.upc.edu/en) and *(ii)* the Database Technologies and Information Management group (**DTIM**, https://www.essi.upc.edu/dtim/). These two subgroups together cover the relevant aspects related to SE and DE that lay the foundations for DOGO4ML.

The rest of the paper is organized as follows. Section 2 and Section 3 present the conceptualization of DOGO4ML and its objectives, respectively. The expected outcomes of the project are detailed in Section 4. Section 5 sketches the relevance of the project for the ML field. Then, Section 6 summarizes the initial results of DOGO4ML. Finally, Section 7 presents the conclusions.

2. DOGO4ML Conceptualization

The main objective of the project is to provide a holistic approach to MLSS engineering aligning its DE needs with SE practices. DOGO4ML proposes a holistic end-to-end framework to develop, operate and govern MLSS and their data. This framework revolves around a new proposal we call the DevDataOps lifecycle, which unifies two software lifecycles: the DevOps lifecycle and the DataOps lifecycle. The DevOps cycle aims to transform the requirements of an MLSS into deployed code (Dev) and get feedback as soon as possible from the end-users (Ops). This can be used to evolve the requirements (including those that apply to the ML models). The DataOps cycle provides support to the data management and analysis processes that characterize MLSS. The DataOps processes are inter-related with those in the Dev phase of the DevOps software cycle, since they produce the required ML models (created through several iterations in the DataOps lifecycle) to be embedded into the ML software components of the MLSS. Further, the DataOps cycle aims to get feedback from the data analysts to continuously improve the data management and analysis processes. A detailed explanation of the conceptualization of both cycles follows.

2.1. The DevOps software cycle

DevOps is a software development and delivery process that produces software from its conceptualization, as well as from the feedback provided by monitors when the software system is in an operational environment. This feedback is then used to maintain and evolve the system. The specificity of MLSS requires continuous context-aware delivery, and feedback to adjust and refine their embedded ML components. Fig. 1 presents the resulting DevOps cycle.

In the Dev phase, a typical requirement engineering process applies both at the system and the ML component levels. At the system level, the requirements include quality requirements specific for the MLSS (e.g., trustworthiness and ethics), extracted from a requirement patterns catalogue based on [2] to be built during the project. At the level of the ML components, requirements also include quality requirements of ML models (e.g., model accuracy and low latency) which are key to identify and process the relevant data for the ML model construction, validation, and operation (see description of DataOps software cycle in Section 2.2).

Going on with the Dev phase, agile practices will be adopted to continuously deliver high-quality MLSS. The definition of reference architectures and best practices (e.g., iterative integration of ML models provided by the DataOps software cycle into ML components), driven by the MLSS quality requirements, will enable rapid MLSS implementation and deployment in small iterations. Automated integration and testing of those systems will reconcile the particularities of both types of components, ML and non-ML (e.g., in terms of uncertainty in the functional validation). Once validated, the MLSS

will be deployed in its contextual operational environment (e.g., in a type of system with high decisional capabilities such as a smart vehicle).
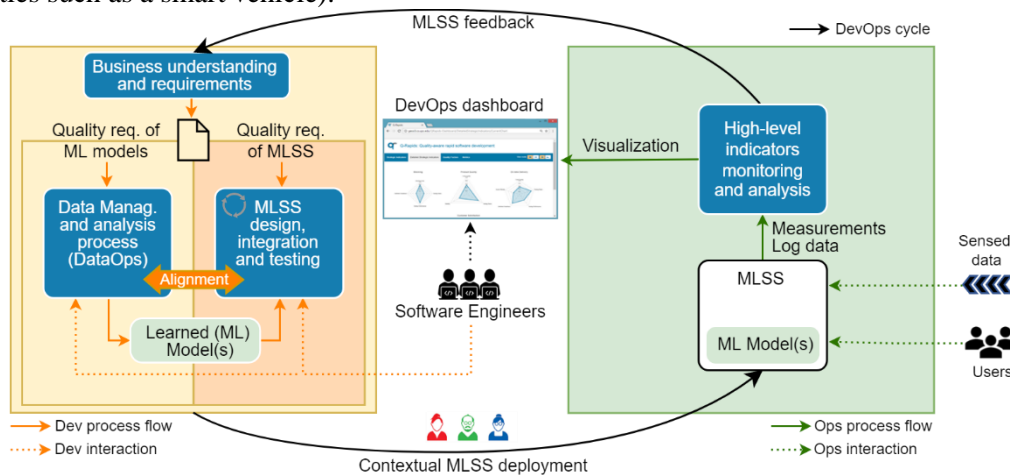
Figure 1: The DevOps cycle for MLSS projects proposed by the DOGO4ML project. A zoom-in of the DataOps cycle is in Figure 2.

During the Ops phase, the MLSS in production interacts with both the user and the environment. For example, a MLSS for a smart vehicle will receive input from the user (e.g., through voice) and a continuous stream of sensed data (e.g., data indicating people crossing the street). Through these interactions and input, the ML models deployed inside the system are able to make predictions (e.g., there is an increased risk of accident).

While the system is in execution, it generates runtime data, mainly in the form of measurements of the system behavior (through monitors) and log files that contain the sequence of time-stamped interactions. This data will be gathered by a module able to analyze it and assess a set of high-level indicators that may refer to MLSS quality requirements (e.g., runtime efficiency, trustworthiness of the system) or other more general aspects (e.g., users' ethical behavior). At this respect, we plan to adapt our previous results in: 1) self-adaptive systems monitoring to the area of MLSS [3] and 2) visualization of high-level indicators and quality requirements in the form of a dashboard [4]. This dashboard, which we call the DevOps dashboard, is an essential aspect of the DevOps lifecycle to generate the needed feedback to impact the Dev cycle, and to close the continuous DevOps loop. Feedback enables the evolution of the MLSS (also including the ML components, by evolving the quality requirements of ML models). Then, the approach starts over again, and the Dev phase uses the feedback to evolve the MLSS. Note that data in operation may thus require revisiting the DataOps lifecycle in another Dev phase.

2.2. The DataOps software cycle

DataOps defines the lifecycle of the data management and analysis processes, characteristic aspects of MLSS related to DE (see Fig. 2). The complexity and iterative nature of these processes require their own software cycle specific for data-related aspects. Additionally, these processes are interdependent with DevOps activities undertaken in the Dev phase. It is thus one of the objectives of the DOGO4ML project to identify and operationalize such dependencies.

Data management processes are responsible to ingest, store, process and prepare data according to the requirements gathered. These processes, common for the whole organization, are carried out by the data management backbone system that serves the data in the form of data views (i.e., datasets generated from the wealth of data ingested ready to be consumed). Then, each project, during its requirement engineering process conducted in the DevOps Dev phase, decides the specific subset of data assets (i.e., data views) required. These data views are the main driver enabling the data analysis processes.
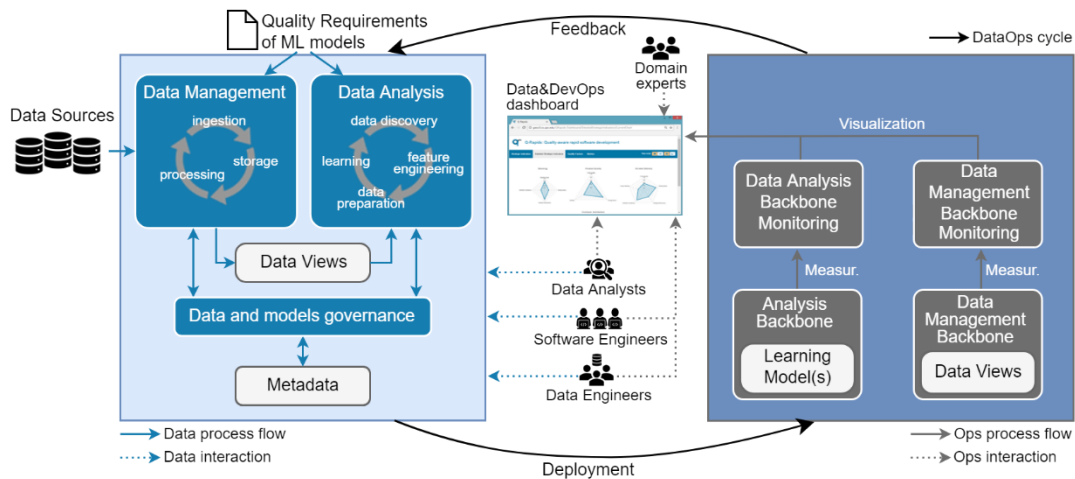
Figure 2: The DataOps cycle for MLSS projects.

Data analysis processes include data discovery (i.e., finding the relevant data assets and requesting the needed data views), feature engineering, data preparation and the model learning. These processes, specific for each project, are carried out by the analysis backbone system that is responsible for learning the models that will be eventually deployed in the Dev phase. Some works frame the data analysis processes into their own lifecycle (e.g., under the concept of MLOps). However, many authors argue that the complete data lifecycle (management and analysis) should be jointly governed within a single unified view [5]. In DOGO4ML we follow this approach and the tasks identified by MLOps are considered in our DataOps lifecycle.

The complexity of the data management and analysis processes requires dedicated data and model governance, embedded in the data governance subsystem. The governance can be achieved by gathering the required metadata to automate, trace, monitor and assess specific requirements for the data management and analysis backbones systems.

Quality requirements of ML models elicited during the Dev phase (e.g., model accuracy), indicators related to learning models (generated during the data analysis processes, such as model appropriateness) and indicators related to data (generated during the data management processes, such as quantifying data bias or query time when accessing the data views) must be monitored during the operation of the data ops cycle and visualized through the DataOps dashboard. Those indicators provide feedback that is key to close the loop with the data cycle. For example, the feedback obtained from monitoring a generated ML model (e.g., its poor accuracy) may require to consider features from another data view, to learn new models or even ingest a new external data source.

2.3. The Holistic software cycle

While the DevOps and DataOps cycles raise significant challenges by themselves, the emerging grand challenge is their combination into an overarching cycle smoothly integrating their different process elements (activities, roles, etc.) into a unique holistic process. We already made a first approximation to the problem in the context of trustworthy autonomous systems [6]. Overall, we envisage three major determinants.

Inter-dependency. Both lifecycles generate a number of inter-dependencies, which, due to the iterative nature of the problem, are not easy to identify, formalize and generalize as to guarantee adaptability to different scenarios.

Context-awareness. We do not aim at defining a universal holistic MLSS lifecycle. Instead, we recognize the fact that different organizations, projects and teams may respond to different context characteristics (e.g., data quality, available human skills, problem size, etc.), and that the MLSS lifecycle needs to be flexible enough as to apply to all of them.

Systematization. To assist software and data engineers in customizing the lifecycle according to context, DOGO4ML proposes a systematic, tool-supported knowledge-based approach that assists them in: *(i)* defining parameterized process fragments (possibly inter-dependent with others) that describe

activities that may take part in the holistic cycle; *(ii)* select the most appropriate process fragments in a particular context, respecting their inter-dependencies; *(iii)* combine them into the holistic process.

Given these determinants, the project will use situational method engineering (SME) [7] as the conceptual framework for defining MLSS lifecycles. In SME, we can define a library of process fragments ("chunks") classified according to some context criteria. We will use our knowledge in context ontologies [8] to define the relevant context criteria in the scope of MLSS. SME supports the composition of such chunks (although the current state of the art does not handle the problem of inter-dependencies), as we have done in previous work (e.g. in the field of software evolution [9]). Tool-support will take the form of a handy web application establishing a conversation with the engineers to proceed with the context criteria elicitation, chunk selection and final composition.

3. Objectives

Based on the aforementioned DOGO4ML conceptualization, we break down the main objective of the project into four objectives:
1. Specify, design and implement a holistic and configurable end-to-end lifecycle for MLSS aligning SE and DE development and operational processes.
2. Specify, design and implement the data-driven Dev phase for MLSS considering quality requirements and architectural aspects.
3. Specify, design and implement the Ops phase increasing users' trust in MLSS by transparently monitoring quality requirements in near real-time.
4. Specify, design, implement and govern the data management and analysis processes for MLSS in the form of a DataOps lifecycle.

4. Expected outcomes

In line with its objectives, the project aims to contribute scientifically to advance the state of the art on the effective and efficient production and continuous evolution of ML models integrated into MLSS. Although ML models and intelligent systems have existed for a long time, the tight integration among models and software from the different perspectives of development, operation, governance and evolution makes this proposal highly innovative.

In particular, this project changes the way in which the interdisciplinary combination of DE and SE proposes the foundation of a new future technology, setting up a baseline (in the form of a proof-of-concept) ready to be matured and transferred to interested actors.

The main assets to be produced in the project that can be individually transferred are:

(1) Process support.
    (1.a) Catalog of SE and DE combinable and customizable process fragments
    (1.b) Catalog of customizable MLSS holitistic processes
    (1.c) Tool support to build the appropriate MLSS process using (1.a) and (1.b)
(2) Development support.
    (2.a) Quality model for MLSS with associated catalogs of requirement patterns
    (2.b) Quality model for ML models
    (2.c) Software reference architecture (SRA) for MLSS
    (2.d) Tool support to instantiate the SRA to a given context considering domain requirements
    (2.e) Tool support for integration testing including ML and non-ML components
(3) Operations support.
    (3.a) Set of tools for the deployment of MLSS with models customized to context
    (3.b) Data ingestion infrastructure for monitoring MLSS specific quality requirements
    (3.c) Set of strategic indicators related to users' trust in MLSS
    (3.d) Strategic dashboard to visualize trust-related indicators
(4) DataOps support.
    (4.a) Set of tools to govern the complete data lifecycle in MLSS.

(4.b) Set of tools to semi-automatically manage data quality aspects.

(4.c) Set of tools to support the automation of the data analysis tasks.

(5) Complete platform. Integrates all the assets above into a single platform.

To promote the dissemination, exploitation and technology transfer of these assets, initial plans have been designed and will be further elaborated as the project progresses.

To foster the dissemination of scientific contributions, the plan targets: *(i)* industrial dissemination, through participation in industry-oriented meetings and dedicated meetings; *(ii)* educational dissemination, incorporating consolidated results at the end of the project into MSc and PhD courses.

Regarding technology transfer and exploitation, the aim is to promote and maximize industry collaborations; capitalization of knowledge and assets in future projects; network growth into new domains. Therefore, an initial business plan has been designed. Such business plan follows the business model canvas approach [10] including the following drivers:

- Create awareness in the partners' ecosystems and beyond (e.g., local networks on the topics of DE and SE standardization through bodies like IREB, see above);
- Deployment of the project ideas into the health, insurance and finance, and open publications domains. Regarding these domains, we plan to conduct empirical studies in companies that expressed their interest in this project to validate the results of the project;
- Get advice from world-leading experts in specific areas to enable the encapsulation of meaningful transferable parts of the project that will ease the project results' transferability and the elaboration of specific supporting tools that will be also integrated to offer an end-to-end support tool.
- The resulting tools from the project will be offered as open source in a public GitHub repository with the appropriate licenses and communicated through adequate channels, as for instance the ReachOut Platform[2], aimed to connect research projects with beta testers and early users on the market.
- To offer a catalog of services of all exploitable resources from the project to foster and facilitate their adoption (installation, maintenance, etc.).

All in all, based on the predicted economic impact of adopting AI/ML from diverse organizations, we expect that the results from the project will contribute to such impact. On the one hand, the World Economic Forum stated that AI/ML are expected to create 133 million new jobs globally by 2022. IDC reported that AI/ML technology spending in Europe for 2019 has increased by 49% over the 2018 figure to reach $5.2 billion. According to a recent survey, AI/ML tools globally are expected to reach US$119 Billion by 2025[3]. Yet, according to market analysis firm McKinsey, "*Most companies are capturing only a fraction of the potential value from data and analytics. [. . .] manufacturing, the public sector, and health care have captured less than 30 percent of the potential value we highlighted five years ago*[4]." This is particularly true in Europe, which is "*lagging behind in embracing the digital and data revolution*[5]." Consequently, any significant advance in the field will have a positive impact not only on economic terms but also socially.

5. Relevance to information science

Interest in the development of artificial intelligence, with particular focus on ML has exploded in the past decade[5]. This has drawn a lot of research activity and investment, enabling ML models to continuously make gains in image recognition, language translation, object recognition, and other applications. The latter has raised the need for incorporating ML models inside conventional software systems, requiring a paradigm shift in terms of how these systems are developed and maintained.

The DOGO4ML project aims to set the foundations for developing software systems that embrace these new possibilities provided by ML. Consequently, the project is set to address different challenges

---

2   https://www.reachout-project.eu

3   https://medium.com/@RWW/how-ai-is-transforming-software-development-ba705e799ca4

4   https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world

5   https://aiindex.stanford.edu/report

that appear as a result of intertwining ML and non-ML components. In particular, since ML models are fed by data, one of the main challenges, with relevance to the information science discipline, is related to data and information management. In this regard, DOGO4ML aims to develop an end-to-end operational data governance framework, identifying and operationalizing the data management lifecycle and analysis processes. Data management processes are responsible to ingest, store, process and prepare data according to the requirements gathered, and then the ready to be consumed data are the main driver enabling the data analysis processes. The complexity and iterative nature of these processes requires a dedicated data and model governance that produces and gathers the required metadata to automate, trace, monitor and assess specific quality requirements. The latter can be related to *(i)* the learned models (e.g., model appropriateness or training time), and *(ii)* the data (e.g., quantifying data bias, data quality or query time when accessing the data views). The first are interpreted and validated by domain experts assessing the quality of the current model, while the second are interpreted and analyzed by the data and software engineers. Finally, once this loop of the data cycle is closed, the resulting learned models are embedded and integrated in the MLSS in the form of an ML component during the Dev phase, as described in Section 2.

6. Current results

Although the DOGO4ML project is at an initial stage, we may report some results. To gain a deep insight about the state-of-the-art related to the project, [11] conducted a systematic mapping study about SE approaches for building, operating, and maintaining AI-based systems. This mapping study provides a consolidated background to tackle the project tasks. Furthermore, [12] developed and scrutinized a generic method that allows to generate pre-processing pipelines, as a step towards automating the data preparation for ML, which in turn is a critical step for the data analysis part.

DOGO4ML will use SME as the conceptual framework for defining MLSS lifecycles (see Section 2.3). As a first result to elaborate this framework, [13] proposed a holistic method, applying SME, to consider data as a new source in requirements elicitation for data-driven systems, as the case of MLSS.

Regarding the development support expected outcomes, DOGO4ML aims at providing a software reference architecture for MLSS and tool support to instantiate this architecture to a given context considering domain requirements. In this sense, an analysis of the impact of design decisions on the achievement of high-accuracy and low resource-consumption in the context of AI mobile applications are provided in [14]. Additionally, in the sentiment analysis domain and as a proof-of-concept, [15] and [16] proposed an architecture able to monitor and analyze the sentiment of tweets shared by end-users.

7. Conclusions

In this paper we have presented the goals and vision of the DOGO4ML project. The expected outcomes and initial results are detailed. So far, all the project tasks are being developed as expected and the first results confirmed that the project is progressing in the right direction. More information is available on the project website, https://dogo4ml.upc.edu/en.

Acknowledgements

References

[1] A.L. Samuel, Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development 44 (1959) 206–226. doi: 10.1147/rd.33.0210.
[2] M. Oriol, S. Martínez-Fernández, W. Behutiye, C. Farré, R. Kozik, P. Seppänen, A. M. Vollmer, P. Rodríguez, X. Franch, S. Aaramaa, A. Abhervé, M. Choraś, J. Partanen, Data-driven and Tool-

supported Elicitation of Quality Requirements in Agile Companies, Software Quality Journal 28 (2020) 931-963. doi: 10.1007/s11219-020-09509-y.

[3] E. Zavala, X. Franch, J. Marco, Adaptive Monitoring: A Systematic Mapping, Information and Software Technology 105 (2019) 161-189. doi: 10.1016/J.INFSOF.2018.08.013.

[4] L. López, M. Manzano, C. Gómez, M. Oriol, C. Farré, X. Franch, S. Martínez-Fernández, A.M. Vollmer, QaSD: A Quality-aware Strategic Dashboard for supporting decision makers in Agile Software Development, Science of Computer Programming 202 (2021). doi: 10.1016/j.scico.2020.102568.

[5] V. Khatri, C.V. Brown, Designing Data Governance. Communications of the ACM 53 (2010) 148-152. doi: 10.1145/1629175.1629210.

[6] S. Martínez-Fernández, X. Franch, A. Jedlitschka, M. Oriol, A. Trendowicz, Developing and Operating Artificial Intelligence Models in Trustworthy Autonomous Systems. In: Cherfi, S., Perini, A., Nurcan, S. (eds) Research Challenges in Information Science. RCIS 2021. Lecture Notes in Business Information Processing, vol 415. Springer, Cham. doi: 10.1007/978-3-030-75018-3_14

[7] B. Henderson, J. Ralyté, P. Ågerfalk, M. Rossi, Situational Method Engineering. Springer, 2014.

[8] O. Cabrera, X. Franch, J. Marco, 3LConOnt: A Three-level Ontology for Context Modelling in Context-aware Computing. Software System and Modelling 18 (2019) 1345–1378. doi: 10.1007/s10270-017-0611-z.

[9] X. Franch, J. Ralyté, A. Perini, A. Abelló, D. Ameller, J. Gorroñogoitia, S. Nadal, M. Oriol, N. Seyff, A. Siena, A. Susi, Situational Approach for the Definition and Tailoring of a Data-Driven Software Evolution Method, in: Proceedings of the Advanced Information Systems Engineering, CAiSE'18, Springer, 2018, LNCS 10816, pp. 603-618. doi: 10.1007/978-3-319-91563-0_37.

[10] A. Osterwalder, Y. Pigneur, Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers. Wiley, 2010.

[11] S. Martínez-Fernández, J. Bogner, X. Franch, M. Oriol, J. Siebert, a. Trendowicz, A.M. Vollmer, S. Wagner, Software Engineering for AI-Based Systems: A Survey, ACM Transactions on Software Engineering and Methodology, Vol. 31, No. 2, Article 37e (2022), 59 pages. doi: 10.1145/3487043

[12] J. Giovanelli, B. Bilalli, A. Abelló, Data pre-processing pipeline generation for AutoETL, Information Systems, In Press, 2021. doi: 10.1016/j.is.2021.101957.

[13] X. Franch, A. Henriksson, J. Ralyté, J. Zdravkovic, Data-Driven Agile Requirements Elicitation through the Lenses of Situational Method Engineering, in: Proceedings of the IEEE 29th IEEE International Requirements Engineering Conference (RE'21), RE@Next! Track, 2018, pp. 402-407. Doi: 10.1109/RE51729.2021.00045.

[14] R. Creus, S. Martínez, X. Franch, Which Design Decisions in AI-enabled Mobile Applications Contribute to Greener AI?, ESEM 2021, URL: https://arxiv.org/abs/2109.15284.

[15] A. de Arriba, M. Oriol, X. Franch, Merging Datasets for Emotion Analysis. An Approach using BETO on Spanish Tweets, 2nd International Workshop on Software Engineering Automation: A Natural Language Perspective, NLP-SEA@ASE'21). doi: 10.1109/ASEW52652.2021.00051.

[16] A. de Arriba, M. Oriol, X. Franch, Applying Transfer Learning to Sentiment Analysis in Social Media, Proceedings of the 5th International Workshop on Crowd-Based Requirements Engineering (CrowdRE'21), 2021, pp. 342-348, 2021. doi: 10.1109/REW53955.2021.00060.