Sophie Jentzsch and Nico Hochgeschwender*

A qualitative study of Machine Learning practices and engineering challenges in Earth Observation

https://doi.org/10.1515/itit-2020-0045
Received October 10, 2020; revised March 4, 2021; accepted June 16, 2021

Abstract: Machine Learning (ML) is ubiquitously on the advance. Like many domains, Earth Observation (EO) also increasingly relies on ML applications, where ML methods are applied to process vast amounts of heterogeneous and continuous data streams to answer socially and environmentally relevant questions. However, developing such ML- based EO systems remains challenging: Development processes and employed workflows are often barely structured and poorly reported. The application of ML methods and techniques is considered to be opaque and the lack of transparency is contradictory to the responsible development of ML-based EO applications. To improve this situation a better understanding of the current practices and engineering-related challenges in developing ML-based EO applications is required. In this paper, we report observations from an exploratory study where five experts shared their view on ML engineering in semi-structured interviews. We analysed these interviews with coding techniques as often applied in the domain of empirical software engineering. The interviews provide informative insights into the practical development of ML applications and reveal several engineering challenges. In addition, interviewees participated in a novel workflow sketching task, which provided a tangible reflection of implicit processes. Overall, the results confirm a gap between theoretical conceptions and real practices in ML development even though workflows were sketched abstractly as textbook-like. The results pave the way for a large-scale investigation on requirements for ML engineering in EO.

Keywords: Machine Learning, Artificial Intelligence, Earth Observation, Process Models

ACM CCS: Computing methodologies → Machine learning, Social and professional topics → Professional topics → Management of computing and information systems → Project and people management → Systems development, Software and its engineering → Software creation and management → Designing software → Requirements analysis, Human-centered computing → Human computer interaction (HCI) → HCI design and evaluation methods → User studies

1 Introduction

Machine Learning (ML) is on the rise and is examined and utilised in many different fields of both research and industry. When it forms the basis of a vast variety of applications, it increasingly becomes a fundamental element of practice and progress. However, the surrounding infrastructure and support resources for the conscious and responsible development of ML systems lag behind. There is a lack of standardised approaches to store, manage and explain data, artefacts and experimental results [16, 12]. The pursuit of new innovative applications and of even higher accuracy scores in existing tasks dominates the field and there are seldom enough resources to carefully plan, document or reflect the executed experiments [17]. This is especially devastating in deep learning, where processes and models are often particularly opaque and black-box-like [12].

To harness the full potential of ML and to enable responsible development of ML-based systems more attention needs to be paid to understanding and supporting the underlying development processes. This includes the demand for efficient and established ways to document provenance information, model characteristics and limitations of a system [8]. To build common standards and to mature supporting systems more effort needs to be spent in software engineering solutions [1].

Likewise, the field of Earth Observation (EO) underwent a paradigm shift from computational to data-driven science, making ML approaches increasingly indispensable [27]. Remote Sensing devices are constantly in operation and produce tons of data. Satellites of the European Copernicus project, for example, produced approximately 6 TB of data every single day in 2016. The amount of data will increase even more through the years and makes domain experts facing the challenges of big data [11]. To this

***Corresponding author: Nico Hochgeschwender,** DLR Institute for Software Technology, Cologne, Germany, e-mail: nico.hochgeschwender@dlr.de, ORCID: https://orcid.org/0000-0003-1306-7880
Sophie Jentzsch, DLR Institute for Software Technology, Cologne, Germany, e-mail: sophie.jentzsch@dlr.de, ORCID: https://orcid.org/0000-0001-6217-8814

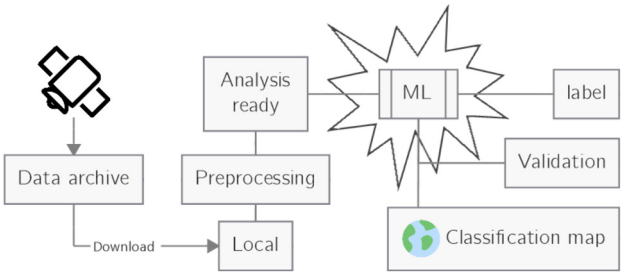

Figure 1: Workflow No. 2 (WF2) – Example ML Workflow sketched by one of the experts in the Think-Aloud task.

end, to ultimately support researchers from EO in structuring ML engineering, and to build appropriate software engineering tools we first aim to empirically understand real circumstances and requirements. While ongoing research mostly focuses on the perception of professional data scientists in general [21, 1], we argue that developers from specific domains are likely to have individual requirements which should be carefully investigated. Moreover, transparency and verifiability are in terms of open science especially worthwhile in a domain with huge scientific drive and impact, as EO.

The present manuscript, reports an exploratory study that aims for the following contributions: First, to explore current development practices. Special attention is dedicated to researchers' perception of ML workflows, as they play a pivotal role in collaboration and exchange within a group. Second, to identify challenges in ML development from the perspective of EO domain experts. Third, to sensitize for critical aspects that need to be considered in similar follow-up investigations. This initial investigation aims to test the expediency of the investigation approach before applying it widely.

To empirically gathers insights into tasks and ML-related challenges semi-structured interviews were conducted. These extended with a novel think-aloud task to capture internalised workflow conceptions of consulted domain experts. This task is hereafter referred to as the workflow sketching task. Reported observations focus on two main research questions:

(RQ1) ML workflows: How do experts perceive their own workflows? How is the expert's conception of the workflow structured? How similar are individual conceptions among different members of the same group? How do the conceptions look like compared to textbook depictions?

(RQ2) Challenges: What do experts perceive to be the biggest challenges in their own work, in EO, and in the general field of ML? What requirements can be derived from that?

The remainder of this article is structured as follows: Section 2 briefly outlines the importance of ML in EO. Both the interviews and the workflow sketching task are described and motivated in Section 3. Results are presented in Section 4, interpreted and discussed in Section 5, and concluded in Section 6.

2 ML for Earth Observation

With technical advances in the field of Earth Observation (EO), the quantity of available data grows exponentially. Nowadays, satellites and research aircraft are constantly in operation to collect a vast amount of heterogeneous types of information about the world and its surface. Recent ML methods enable the processing and analysis of these vast amounts of heterogeneous and continuous data streams to find general patterns and infer hidden correlations. This property makes ML an essential pillar of recent EO breakthroughs. Beyond scientific advances, applications possess the potential to contribute to the global welfare and to reform essential aspects of our society as there are several humanitarian applications [15] and environmental aimed applications [23] of ML and remote sensing data. Some notable examples are, to name a few, tracking the retreat of snow and ice-cover [24, 13, 4] to measure the impact of global warming or exploiting satellite data to analyse refugee camp behaviours to forecast streams of refugees [6]. These are only a few of many noteworthy examples of the exploitation of EO applications using ML techniques. In this regard, deep learning (DL) approaches are widely applied and discussed [27]. As Zhang et al. put it: *"[deep learning] is actually everywhere in [remote sensing] data analysis"* [26]. Besides national and international aerospace organisations holding an exceptional leading role and possess wide ranges of domain-specific knowledge, there are already commercial players capable to compete in the run for new ML advancement. These are often well-experienced in economically efficient and product-orientated software development. To support traditional EO institutes in keeping up even better with the high pace of the ML development field, it is necessary to understand the situation and researchers' perception regarding the shift to data-driven approaches.

3 Methodology

In this section, we describe the general design and analysis procedure of conducted interviews. The uninformed participants, who are mostly referred to as experts, were

defined to be EO researchers who work on ML approaches. The expert interviews were enriched with a novel implementation of a think-aloud task that was designed from scratch and provides even more unbiased insights into the experts' implicit mental conceptions of ML workflows. The interview transcriptions were analysed with coding techniques. Both tasks and the analysis are described in the following three sub-sections.

3.1 Semi-structured interviews

Conducting semi-structured expert interviews is a well-established procedure, originally from psychology and social science, and proven for the exploratory examination of requirements in both the contexts of software engineering [20, 14] and ML [3, 7]. Compared to fully structured interviews there is some flexibility regarding the course of a conversation, which allows the experts to express their very own perception without being primed by too explicit questions. An interview guideline was predetermined to help the interviewer to keep an overview of covered core topics and to associate the experts' statements with research questions afterwards. It also ensures comparability and methodological validity.

The interview covered about 30 questions in total, which were subdivided into seven main blocks. For each of these blocks, there were defined lead questions that address the core information, and follow-up questions that were anticipated to be useful in specific situations. A comprehensive list of questions can be found in the appendix (Sec. 7.1). The interview was divided into the following seven main parts:

(1) Technical Background – Basic information as ML-related education, previous work experience, and the current research topic.

(2) Current work – Research topic and usual everyday task. Particularly, it was aimed to understand on what level (theoretically, application orientated, development orientated), to what extent, and by which means researchers work with or on ML processes.

(3) Individual workflow – Conduction of think-aloud task (Sec. 3.2), followed by prepared questions regarding specific characteristics of the individual workflows.

(4) ML in own work – Details of particular steps, e. g. the effort of time and work, collaboration, challenges and troubleshooting.

(5) The Situation in ML – Shift of focus from an individual to a more general context, namely experts' view on best practices and state-of-the-art in a more general context.

(6) Self-assessment – Experts' perception of their own skills and obstacles in the context of ML.

(7) Closing questions – Explicit questions, e. g. concerning engineering challenges, were placed at the very end to avoid priming. Finally, participants had the opportunity to further comment and pose questions.

3.2 Think-aloud workflow sketching

Several different design and implementation decisions need to be taken on the way to create ML-based systems. That may include among others handling a complex data basis, different data preprocessing and data cleaning steps, model selection, model implementation, hyperparameter settings, model deployment, and mastery of experimental results. Even though conceptual steps can be mostly similar, the entire process usually requires a considerable amount of exploration. These typical steps and how they are logically and chronologically connected in reality are what we refer to as workflow. We aim to understand to which extent experts are involved in the different steps and how much they are aware of their own workflow. This desired insight is particularly conceptual and complex. Implicit conceptions are not always well-accessible with explicit questions, especially when they include complex dependencies and subtle differences. To capture experts' perception of ML workflows as accurate as possible the interview was widened by a special element, inspired by think-aloud tasks [19]. Participants were asked to sketch their workflow on a blank piece of paper. No other information was provided and instructions were formulated as open as possible to not imply any direction. There were no specifications regarding the included elements, layout, perspective, etc. The drawings were evolved very spontaneously and do not imply completeness or accuracy. They are intended to reflect the individual perception of experts and the general structure of cognitive conception. Within the ML community, there is a broad understanding of what ML workflows look like in reality. Numerous practical blog posts and textbooks aim to make these abstract conceptions tangible (e. g. [22, 20, 5, 1]). Of course, there are differences in the variety of depictions. There is not one single implementation that is generally valid in all contexts. However, certain characteristics can repeatedly be found in exemplary workflows, which can be utilised to describe similarities and differences. We derived the following set of reference characteristics:

Sequential: The overall framework has sequential chronology like a box-and-arrow format. It is not just a collection of unsorted bullet points.

Iterative: There are loops, cyclic dependencies or representations of decisions included.

Generic: It is an abstract depiction that is generally valid, also for other people and different tasks.

Closed: The workflow consists of one overall system, where all elements are embedded.

View: What are the perspective (e. g. developer- or data-centred) and the scope of workflow? Are some parts more prominently represented than others?

Accessible: Is it easy for the experts to withdraw the process and to put it down on paper?

3.3 Information processing and analysis

Since it is explicitly not intended to criticise or exposure the work of contributing experts, every effort was made to handle intimate information sensitively and to consider ethical reservations [18]. Each interview took about one hour, leading to 320 minutes of voice record raw data. Audio recordings were taken during the interview and transcribed and anonymised immediately afterwards. Personal details and linguistic idiosyncrasies that do not convey contextual information were removed conscientiously. Thus, included citations may not reproduce the very exact wording of interviewees in favour of intelligibility and integrity of anonymity. For the same reasons, workflows were not depicted with original handwriting but reconstructed neatly. The analysis of data was conducted through coding in multiple extensive iterations of cleaning, reading and annotating as described (see [2]). First, categories were determined deductively on basis of research questions. For each interview, relevant statements that were spread all over the conversation were assigned to the categories. Additionally, exploratory coding was applied based on the challenges that were brought up by the participants. For instance, when a challenge was discussed by one expert it has been cross-checked, whether and to what extent the others mentioned that or something similar as well. This way, the weight of a certain issue could be derived from the frequency, extensiveness and emphasis that it was discussed with.

In the think-aloud task, the complete thoughts were uttered verbally, while only keywords were noted down by the participants, which alone can be ambiguous. Therefore, verbally uttered explanations in the transcriptions were linked to the corresponding elements in the sketches. This enabled a sophisticated interpretation of experts' expositions for the analysis of workflows. Also, the chronological order in which experts noted particular elements was relevant, e. g. for the evaluation of the characteristics *sequentiality* and *accessibility*, and could be obtained from transcriptions. With this preparatory work, it was possible to compare sketches by the defined characteristics. Finally, it was examined to what extent the (3) individual workflow of each participant matches the oral description of their work [(2) and (4)]. On basis of the contextualised workflows and the composed responses to the particular categories, it was possible to derive answers to the main research questions.

4 Results

Results are roughly structured according to the research questions that are stated in the Introduction: The workflows (RQ1) are, first, characterised by defined criteria of textbook-like depictions (Sec. 4.1) and, second, by exploratory developed criteria (Sec. 4.2). Subsequently, the most prominently discussed engineering-related challenges (RQ2) are discussed in Sec. 4.3. First of all, experts are briefly characterised:

The experts were researchers (one female, four male) between 24 and 31 years old with different native countries. The duration of employment in the current research group varied considerably, as two of the experts just started to work in that position (less than 6 months) and the other three were members in that group for some years already (three years or more). Yet, looking at the total job experience in that scientific field, all participants appeared to be on a fairly equivalent level.

Experts graduated in technical subjects that were all directly related to EO or deep learning in any way. Backgrounds were still fairly individual, as some of the experts are related more to computer science and others rather stem from classical Earth Observation. All experts stated that their university education covered ML approaches to a certain extent, even though the focuses on mathematical basics were unequally strong. All five participants explicitly define ML to be (one of) their core area(s) themselves during the interview.

4.1 Workflows by criteria

The experts' workflow assessments were captured in the think-aloud task. For analysis, oral explanations were linked to the sketched illustrations and finally categorised utilizing pre-defined characteristics. Observations regarding each characteristic are discussed in the following, accompanied by an overview of interviews and categories in Table 1. One of the exemplary sketches is displayed in Figure 1, the others can be found in the appendix.

Table 1: Characteristics of Workflow Sketches – View is either focused on the Researcher's perspective (R), the Task(s) (T), or Data (D).

Criteria	WF1	WF2	WF3	WF4	WF5
Sequential	√	√		√	
Iterative	√	(√)	(√)	√	√
Generic	√	√		√	
Closed	√	√		√	
View	R	D	T	R	T

In general, results demonstrate how experts implicitly understood the task differently under the same instructions: The authors of WF3 and WF5 had a strong focus on their individual tasks and endeavoured to explain them vividly, whereas the other three experts intuitively took a step back to describe the bigger picture in a textbook-like manner.

Sequentiality

WF1, WF2 and WF4 meet the expectations of a box-and-arrow depiction of processes. Steps and dependencies are clearly indicated by arrows. Also, WF3 and WF5 include successive steps, but the form of presentation was more like listed bullet points.

Iterations

None of the workflows was outstandingly comprehensive regarding iterations. WF1 and WF4 include one loop illustrated with arrows. WF5 does not make use of arrows in the first place, but dependencies are indicated by if-statements. WF2 and WF3 do not explicitly show any iterations, but they were implied during all interviews, where participants spoke of validation. Thus, the reason why experts did not include iterations in their explanations seems to be that they considered these dependencies to be clear already: "*Of course we have here in the machine learning some validation.*"

Generic

Even though WF1, WF2 and WF4 capture different views they are all equally generic. Included steps are formulated in an unspecific manner like "adaption" and "implementation". One expert stated: "*I just want to let you get the big picture*". Experts were very aware that their drawings quite simplify real conditions: "*If you really want to dig into details, I can do that. I mean for each of the steps there are a lot of details. Especially in the machine learning part.*" In contrast, WF3 and WF5 focused a lot on the experts' current task, as reflected in the following statement: "*The drawing describes my own workflow, not any standard. They are quite different and especially for the software part it is really individual*". These workflows include more individual steps.

Closed

WF1, WF2 and WF4 each follow one overall framework, where all sketched entities are integrated. Compared to that, WF3 and WF5 are structured differently: Drawings consist of multiple detached sub-elements that are used to explain certain details. Side stages for instance broached the issues of concrete methodological approaches or particular tasks within their work. Therefore, identifying the overall structure was more complex in analysis and the corresponding figures deviate more from the original sketches.

View

WF2 has a very strong focus on the data part of ML, as half of the included steps are concerned with data acquisition, preprocessing and labelling. The core ML part, where the learning happens, is still displayed as a black-box step. All other workflows include data as well, but not as prominently, as they only mention it once each. WF1 and WF4 also display generic workflows, but from a researcher-specific perspective as they include steps like "literature review" and "writing paper". Still, the focus is of course on ML-related work. WF3 and WF5 also describe the ML development from the researcher's perspective, but in a narrower scope related to individual problems. In Table 1, the different views are presented as "D" for data orientated, "R" for researchers' perspective, and "T" for task-oriented.

Accessibility

When participants were asked to draw their workflow, all of them started immediately to bring their thoughts to the paper and explained their steps confidently. This task did not seem to be difficult or intimidating for any of the experts. Minor differences could be observed regarding stringency and chronological structure of responses: Authors of WF1, WF2 and WF4 were quite determined and sketched the final version straightforwardly and without much rethinking, meaning they start their explanation with the first step and finish with the last one in general. One of the experts explicitly stated that (s)he does not draw this for the first time: "*I draw this in my paper (laughs)*". Also, they draw nothing besides this one workflow. In contrast, the processes of sketching WF3 and WF5 were embedded in the conversation and thus more dynamic: Experts went

more into detail at certain points, jumped back and forth in their explanations and were more specific in their elaborations, which is also connected to the level of abstraction. In these cases, corresponding Figures (2 and 5) do not show the whole amount of information but only those strictly related to the actual workflow.

All experts agreed that the sketched workflow is clear to them, but not easy to show and to explain to others, e. g.: "*It is clear, but it is not easy to clearly show it, because it is not linear.*"

4.2 Workflow by categories

The exploratory study design left room for the interviewees to express individual and specific thoughts and to steer the course of conversation. Thus, besides the analysis of pre-defined workflows criteria in 4.1, categories have been identified exploratorily in the transcriptions of experts' explanation. The most prominent and significant categories were *ML as Black-box*, *intuition and experience in ML development*, *troubleshooting* and *Data processing*.

Is ML a black-box?

A common conception of ML models, mostly deep learning models, is that they have black-box characteristics [12], where input and output can be observed, but particular processes are hidden underneath an opaque surface. This issue has been addressed explicitly by four of the experts. All of them share the same opinion, which can be summarised as follows: First, they differentiated between ML techniques, as some of the classical approaches are comparatively transparent anyway. The back-box analogy only refers to the opaquer ones, as deep learning approaches. "*For some traditional machine learning algorithms, we can almost totally understand them. We can clearly see every step. But for deep networks, it is a little bit difficult to fully understand them.*" Second, it depends on the perspective. Outsiders might perceive deep learning as a black-box, while researchers and developers from that field do not necessarily. "*I think it depends on different people looking at this. There are definitely people who don't know what's happening inside.*" Third, technological capacities to "open the black-box" change over time. There is a lot of research effort and processes are much better understood as they were in earlier years. Experts are optimistic that there will be further breakthroughs on that. "*In the future, it is maybe possible to understand a DNN fully.*" None of the experts agreed that ML is a total black-box. However, they also do not entirely discard the black-box idea. One expert described this condition as "grey-box".

Intuition and experience in ML development

Another common conception is that ML development processes are largely driven by researchers' intuition and experience and rather based on trial-an error [8, 16] than on reasonable and traceable development standards. All experts explicitly agreed to that assumption. Remarkably, participants repeatedly touch this issue in statements on other topics, as "*These architectures are based more on intuitions*" or "*We need to decide with our mathematical experience.*" Some statements are even more explicit: "*I am usually tuning hyperparameters based on my own experience. I mean there are no equations to calculate the best learning rate.*" One of the experts continuously describes the situation in ML as "*a big mess*". The high incidents of such comments indicated the urgency of this issue both implicitly and explicitly. It could be observed that experts indeed perceive intuition and experience to play a role in scientific ML development.

Evaluation and troubleshooting

An evidently relevant question is how to evaluate the current state of work, e. g. define the point at which the model is good enough, and how to handle problems. Only one of five experts even mentioned a scenario with clearly defined requirements, indicating this to be an exceptional case. The experts prominently referred to accuracy measures and cross-validation. Yet, when digging deeper it always broke down to the comparison with literature references and state-of-the-art: "*I need to evaluate the model on the test data set, where I obtain the accuracy as the final result. Then I compare this result with all the other results.*" and "*The baseline is what you see in the literature for the task.*" Obviously, there are not many noteworthy alternative options of quality assurance. "*You always want to get better than before, but nobody knows how good it can work.*" Likewise, it seems to be tough to define a goal or end state: "*In the end – Yeah, when is the ending? It is hard to say. You can always try to get better.*"

Experts also reported difficulties that they face in their work, but could not name any concrete problem-solving strategies. Discussed measures were for instance web search, or asking colleagues for help: "*I usually google it first to see if I can get some practical results. If not, I talk to other people, who have a lot of experience.*" None of the experts concretely mentioned any best practice troubleshooting strategy or agreed processes documentation, which suggest that these kinds of solutions are rare.

The role of data

Everyone in ML must care about data in some way. This fact is also substantiated by the observation that *data* as a keyword is included in all workflows. Experts considered data processing to be among the most critical workflow steps. One of them explained: "*Some people say the analysis of data, but so far this is not the case for me*" while another one stated: "*The nature of the data is still really critical.*" Three of five participants emphasized the complexity of data-related tasks: "*The difficulty in handling the data is to figure out how to process it and how to save the data efficiently. This is hard as the amount of data is really huge*". However, researchers are dealing with data to different extents, as some experts focus more on methodological ML approaches. Two of them even described that they get data and just apply their methods to that. We observed that the nature and availability of included data determines individual processes a lot and is a central factor for requirements in ML development. This needs to be considered in future investigation iterations, e. g. by differentiating between data scientists and ML algorithm developer.

4.3 Challenges by category

In the context of daily processes and tasks, experts extensively discussed individual challenges. They reported struggling mostly with engineering-related tasks and not always feeling confident about that part of ML. "*For me, there are a lot of challenges from the engineering point of view*", is just one of many statements. This became especially evident in contrast to scientific challenges that were described as rather general issues. By contrast, experts seem to attribute engineering difficulties to their personal skills such as "*I am not a great coder and I do not have a background in computer science or software development*".

While scientific and engineering aspects have been equally discussed in the interviews, this report focuses on engineering issues. They give useful indications regarding development processes and enable us to derive concrete requirements. Scientific challenges are mostly recognised within the domain already and thus bear little novelty. Scientific categories were data sparsity in weakly or unsupervised learning, the opacity of DNNs, hyperparameter tuning with AutoML, and uncertainty of predictions.

Three main engineering categories could be identified, which are concerned with dependencies, experiments and models, and hyperparameters. These main categories and mentioned solution approaches are described in the following:

Dependencies in soft- and hardware

There is rapid progress in the field of ML and the updates of tools and packages are accordingly frequent and substantial. Experts reported struggling with keeping their ML pipeline up-to-date and aligned. The experts reported huge transfer cost to convey existing code from one setting to another, e. g. from Caffe to TensorFlow. This does not only restrict to the overall framework but also includes updates of side packages or hardware changes. One of the experts explained it as follows: "*The problem with TensorFlow is that [...] everything needs to be perfectly aligned. And if anything goes up or down, it doesn't work*". Even if the process might be clear in theory there can still be a huge overhead in alignment and maintenance of system settings.

Managing experiments and model versions

The structuring of experiments includes keeping track of different models, data and the provenance of information. The experts explicitly pointed to this issue and uttered concerns to different points of the interviews. They reported to struggle with reasonably organising ML projects from scratch and maintaining and nurturing the system when it grows. This especially pertains to bigger projects with various developers involved. "*The problem with machine learning is that there are always too many choices. [...] It is really cumbersome.*" Experts talked a lot about the " *changing anything changes everything*" characteristic of ML development [17]: "*I start training it and then I need to change something. Then I change something small in the code and I run it. So, I have a previous experiment and a new one. And in the end, I have maybe 50 experiments. Now [the challenge is] to see what were the changes that I did on experiment X.*" One described main challenge is to keep track of what change caused which effect on the outcomes.

Hyperparameter setting and tuning

"*Hyperparameters are always a big mess.*", which is a well-known problem in ML model training. This was also discussed a lot in the interviews and does not only include optimisation and documentation of applied parameters, but ideally a reasonable justification of the settings in the first place. According to the experts, there is nothing like guidelines or agreed best practices that tell how to handle hyperparameters, which makes it especially subject to the experience of developers. They stated that they use given default values from the literature if there are any, even if they would not entirely comprehend them: "*I think the default value is there for a reason.*" However, one expert explicitly criticised that there are too many people out there

who "*don't care what's happening inside. They just put the data in and use the results.*"

Discussed strategies on hyperparameter setting include finding the best or at least acceptable parameters by intuition. While some of the experts seemed to be versed in that regard, others are not that concerned with that: "*Hyperparameter tuning is completely different. This is a very big topic and I don't do that at the moment.*"

The issue of hyperparameter setting was discussed both from an engineering and a scientific view. Again, participants were way more optimistic about the scientific part: "*Auto-ML is intensively working on it and I think, this could be solved quite easily*".

Solution approaches

Experts also speculated how existing software engineering solutions could be tailored to their specific engineering challenges. They discussed the deployment of practical guidelines: "*If there were some established references or guidelines that one should follow, maybe that would help.*" These should be generally valid but still need to be task-specific to a certain extent: "*Something that is giving steps you should follow for each case and tells how the test performances are. Where you have a metric of how good your model will be before you even built it.*"

Above that, experts even described a concrete software tool to organise experiments and answer questions as "*what part of the data set did I use*", or "*what is the difference from one experiment to another.*" Still, they were little optimistic about the feasibility: "*I don't know if that is possible.*" Further mentioned aspects that improve the ML development decisively are appropriately established repository management systems and the strong open source community in computer science: "*There is an online community and people post a lot of the problems*".

5 Discussion

In the workflow sketching task, we observed that interviewed ML researchers have a strong mental conception of processes. With defined characteristics (Tab. 1) experts' implementations can be classified into two groups: WF1, WF2 and WF4 are pretty close to the hypothetical textbook-like workflow depictions, which is generic, iterative and closed. They did not draw their *own* workflow that much but stuck pretty much to the abstract general conception. In contrast, the authors of WF3 and WF5 explained more specific procedures in their current scope of tasks, which led to a more detailed and realistic, but also less clear presentation.

Interestingly, response patterns in the sketching tasks correlate with the outstanding heterogeneity of employment durations. WF3 and WF5 stem from the two recently employed group members, while the former more abstract depictions stem from the experienced group members. This differentiation is not reflected in interview responses, where all experts gave an equally knowledgeable impression in both the domain of EO and ML approaches. There is a considerable gap between theoretical ML workflow conceptions and the practical development of ML systems in earth observation. With increasing duration of affiliation and experience in a certain project, the experts' conception of ML workflows might approximate to the abstract textbook-like version. Apparently, experts learn to bridge that gap by viewing and explaining their own complex work in a similarly abstract manner. All experts agreed that real-world ML development is much messier and more unclear as the conceptualised presentations give a hint of, as commented by one expert: "*You have to learn it the hard way*".

Deduced workflow-related categories emphasize the necessity of software engineering tools to facilitate responsible ML development: Experts discussed the black-box characteristic of ML models, the experience-driven development and the shortage of troubleshooting approaches. In Sec. 4.3, three main categories of engineering challenges could be identified, which constitutes a strong basis for subsequent research. These categories are **Dependencies in Soft- and Hardware**, **Managing Experiments and Model Versions**, and **Hyperparameter Setting and Tuning**.

To ultimately supply adequate software engineering tools for ML, a larger body of empirical knowledge and concrete requirements are needed (as in [9]). Similar investigations need to be conducted with additional participants, including a more diverse population and different research institutes, to see to what extent the present observations are representative. Further, it would be interesting to observe researchers of the same institute in a long-term study as the paradigm shift in EO still lasts and it can be expected that involved researchers will be increasingly versed. The striking role of data processing needs to be considered in further research iterations, e. g. by explicitly comparing data-scientists' perceptions to algorithm developers' perceptions. The general approach of combining semi-structured interviews with the think-aloud workflow sketching task turned out to be suitable. Interviews can further be sharpened. The think-aloud sketching task was newly implemented and thus needs to be evaluated even

further. Still, it has proven to be a promising approach for capturing deeper insights into internalised conceptions.

It is proven and evident practice to exploratory identify perceptions, requirements and challenges in qualitative research and subsequently derive less flexible items from that for a qualitative large-scale survey (e. g. by Kim et al. [9, 10]). Accordingly, the next step should be to conduct such a qualitative study to substantiate and further refine the present findings.

6 Conclusion

With the shift from computational to data-driven science [27] Earth Observation (EO) increasingly relies on Machine Learning (ML) approaches to process and analyse vast amounts of heterogeneous and continuous data streams. Simultaneously, the engineering aspects of ML become more demanding. Established software engineering measures cannot straightforwardly be tailored to that [1, 8]. To meet the specific requirements of EO researchers, we seek to understand (RQ1) what their ML workflows look like and to what extent they reflect textbook-like depictions and (RQ2) what they perceive to be the main challenge in ML development. In this initial study, the experimental design of semi-structured expert interviews in combination with a novel implementation of the think-aloud task was tested with five domain experts. Results were analysed by predefined categories and with coding.

While sketched workflows were pretty textbook-like, real processes indeed were described as messier and more intuitive. Participants do have a clear conception of abstract workflows but tend to struggle with a lack of structure in ML engineering. Most interestingly, in this sample group participants with long employment duration sketched their own workflow pretty clear and textbook-like, while members that were new in the group tended to describe their individual processes in a realistic and task-oriented way. This discrepancy could not be found in the experts' explicit explanations regarding their everyday work. This observation suggests that development processes do not get less opaque, but developers learn to map their tasks to abstract conceptions and to present their work properly with time. Furthermore, experts did not view themselves as core software engineers. While they were pretty optimistic about scientific breakthroughs, they perceived the engineering part to be more challenging. Three main categories of engineering challenges could be identified and serve as a basis for further iterations of investigation. Public research institutes need to be able to keep up with emerging commercial players and to maintain domain knowledge in future advances. To this end, further insights are needed for efficiently supporting researchers from different domains in ML engineering. There are already some proposed tools to support researchers in ML engineering (e. g. [25]). It remains to be seen how such tools can efficiently be integrated into experts' daily work and to which extent they remedy identified issues. A sound body of empirical knowledge regarding specific requirements has to be built to examine their suitability.

Acknowledgment: We would like to thank our colleagues from the DLR Earth Observation Center, who supported us in conducting this investigation. Special thanks also to the interviewed experts, who frankly shared their individual experiences with us.

Funding: This work was supported by the DLR "KI Softwaretechnik für die Erdbeobachtung" (AI software technology for earth observation) project.

References

1. Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., and Zimmermann, T. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering* IEEE, pp. 291–300.
2. Cachia, M., and Millward, L. The telephone medium and semi-structured interviews: a complementary fit. *Qualitative Research in Organizations and Management: An International Journal* (2011).
3. de Souza Nascimento, E., Ahmed, I., Oliveira, E., Palheta, M. P., Steinmacher, I., and Conte, T. Understanding development process of machine learning systems: Challenges and solutions. In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)* (2019), IEEE, pp. 1–6.
4. Fischer, G. *Modeling of Subsurface Scattering from Ice Sheets for Pol-InSAR Applications*. PhD thesis, ETH Zurich, 2019.
5. Ge, Z., Song, Z., Ding, S. X., and Huang, B. Data mining and analytics in the process industry: The role of machine learning. *Ieee Access 5* (2017), 20590–20616.
6. Hassan, M. M., Smith, A. C., Walker, K., Rahman, M. K., and Southworth, J. Rohingya refugee crisis and forest cover change in teknaf, bangladesh. *Remote Sensing 10*, 5 (2018), 689.
7. Hill, C., Bellamy, R., Erickson, T., and Burnett, M. Trials and tribulations of developers of intelligent systems: A field study. In *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (2016), IEEE, pp. 162–170.

8. Jentzsch, S. F., and Hochgeschwender, N. Don't forget your roots! using provenance data for transparent and explainable development of machine learning models. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering Workshop (ASEW)* (2019), IEEE, pp. 37–40.
9. Kim, M., Zimmermann, T., DeLine, R., and Begel, A. The emerging role of data scientists on software development teams. In *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)* (2016), IEEE, pp. 96–107.
10. Kim, M., Zimmermann, T., DeLine, R., and Begel, A. Data scientists in software teams: State of the art and challenges. *IEEE Transactions on Software Engineering 44*, 11 (2017), 1024–1038.
11. Koubarakis, M., Bereta, K., Bilidas, D., Giannousis, K., Ioannidis, T., Pantazi, D.-A., Stamoulis, G., Haridi, S., Vlassov, V., Bruzzone, L., et al. From copernicus big data to extreme earth analytics. *Open Proceedings* (2019), 690–693.
12. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications 10*, 1 (2019), 1–8.
13. Parrella, G., Hajnsek, I., and Papathanassiou, K. Estimation of snow and firn properties by means of multi-angular polarimetric sar measurements.
14. Phillips, S., Zimmermann, T., and Bird, C. Understanding and improving software build teams. In *Proceedings of the 36th international conference on software engineering* (2014), pp. 735–744.
15. Quinn, J. A., Nyhan, M. M., Navarro, C., Coluccia, D., Bromley, L., and Luengo-Oroz, M. Humanitarian applications of machine learning with remote-sensing data: review and case study in refugee settlement mapping. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376*, 2128 (2018), 20170363.
16. Schelter, S., Biessmann, F., Januschowski, T., Salinas, D., Seufert, S., Szarvas, G., Vartak, M., Madden, S., Miao, H., Deshpande, A., et al. On challenges in machine learning model management. *IEEE Data Eng. Bull. 41*, 4 (2018), 5–15.
17. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., and Young, M. Machine learning: The high interest credit card of technical debt.
18. Strandberg, P. E. Ethical interviews in software engineering. In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)* (2019), IEEE, pp. 1–11.
19. Van Someren, M., Barnard, Y., and Sandberg, J. *The think aloud method: a practical approach to modelling cognitive*. Citeseer, 1994.
20. Vianna, A., Ferreira, W., and Gama, K. An exploratory study of how specialists deal with testing in data stream processing applications. In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)* (2019), IEEE, pp. 1–6.
21. Vogelsang, A., and Borg, M. Requirements engineering for machine learning: Perspectives from data scientists. In *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)* (2019), IEEE, pp. 245–251.
22. Wang, M., Cui, Y., Wang, X., Xiao, S., and Jiang, J. Machine learning for networking: Workflow, advances and opportunities. *Ieee Network 32*, 2 (2017), 92–99.
23. Wulder, M. A., Loveland, T. R., Roy, D. P., Crawford, C. J., Masek, J. G., Woodcock, C. E., Allen, R. G., Anderson, M. C., Belward, A. S., Cohen, W. B., et al. Remote sensing of environment: Current status of landsat program, science, and applications.
24. Yang, J., Gong, P., Fu, R., Zhang, M., Chen, J., Liang, S., Xu, B., Shi, J., and Dickinson, R. The role of satellite remote sensing in climate change studies. *Nature climate change 3*, 10 (2013), 875–883.
25. Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M., et al. Accelerating the machine learning lifecycle with mlflow. *IEEE Data Eng. Bull. 41*, 4 (2018), 39–45.
26. Zhang, L., Zhang, L., and Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine 4*, 2 (2016), 22–40.
27. Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., and Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine 5*, 4 (2017), 8–36.

Bionotes

Sophie Jentzsch
DLR Institute for Software Technology,
Cologne, Germany
sophie.jentzsch@dlr.de

Sophie Jentzsch graduated in the intersection of Computer Science and Psychology at Technische Universität Darmstadt in 2018. She is currently PhD student and research associate at the Institute for Software Technology of the German Aerospace Center (DLR) in Cologne. The main focus of her research is on human biases in Machine Learning based systems and Explainable Artificial Intelligence (XAI).

Nico Hochgeschwender
DLR Institute for Software Technology,
Cologne, Germany
nico.hochgeschwender@dlr.de

Nico Hochgeschwender is Professor for Autonomous Systems at Bonn-Rhein-Sieg University, Sankt Augustin (Germany) and associated research scientist at the Institute for Software Technology of the German Aerospace Center (DLR) in Cologne. The main focus of his research is on assuring trustworthiness of autonomous and learning enabled systems.

7 Supplementary material

This section provides supplementary material regarding the defined interview guideline (Section 7.1) and the sketched workflows (Section 7.2).

7.1 Interview questions

As described above, the conducted interview was semi-structured. That is, there were defined leading questions, but they were only deployed according to the expert's explanation. Thus, in each section, the first question served as an opening question. All subsequent questions were then organised spontaneously depending on the context and only asked if not already covered by the expert. Further questions were defined to support the expert, in case the initial questions were not clear or misleading. In the following, the main predefined questions for each of the seven sections are listed.

Technical background

Basic information as ML related education, previous work experience, and the current research topic.

- What kind of University degree do you have? What was the focus/ your main subjects? *Is it a Computer Science/ mathematical/ statistical background, or something different?*
- To what extent and how did you learn about theoretical basics and ML techniques?
- Which working group are you associated with currently? What is your position within the working group?
- Please describe your employment background/ career up to now/ previous professional experience.
- What is the title of your current topic? What is the exact name of your current research field?

Current work

Research topic and usual everyday task. Particularly, it was aimed to understand on what level (theoretically, application orientated, development orientated) and to what extent and by which means researchers work with or on ML processes.

- What does your everyday work look like (open description)?
- What projects do you currently work on? (if a PhD student also: What is the topic of your doctoral thesis?
- How is your work connected to ML? What role does ML play in work?
- Do you consider ML methods to be the focus of your work or is it rather a tool to achieve something else?

Further inquiries if needed: Which ML techniques do you use? Which programming language, frameworks, or IDEs do you work with? Which parts do you (have to) implement manually? Do you also use alternative techniques, or have you considered alternative implementations?

Individual workflow

Conduction of think-aloud task (see Sec. 3.2), followed by prepared questions regarding specific characteristics of the individual workflows.

Please sketch your workflow and explain what you are drawing verbally.

- Can you subdivide your work somehow?
- How would you outline the general process?
- How easy is it for you to sketch that? Is some part particularly difficult to sketch? (after the expert finished the sketch)

Depending on the expert's response further questions are needed, for instance:

- What would you do then?
- How would you know that?
- How do you define this phase to be completed?
- What are the individual steps of that?
- What dependencies/ iterations are there? How to handle feedback?
- What do you do when the outcome is not as expected?

ML in own work

Details of particular steps, e. g. effort of time and work, collaboration, challenges and troubleshooting.

- Opening question: Does the drawing reflect your work well?
- Are all phases equally in the focus of your work?
- How labour-intensive/time-consuming do you perceive the individual phases to be? Which phases are more intuitive/ cognitively demanding/?
- Do you work on them alone or in collaboration with others? Is the cooperation cooperative or sequential? Where are the points of contact?
- In which phases do you see the biggest sources of problems?
- Which of the phases are particularly relevant for the predictions of the system in reality?
- How do you deal with them when you encounter problems in development? To whom do you report?

The situation in ML

Shift of focus from individual to a more general context, namely experts' view on best practices and state-of-the-art in a more general context.

- Would the sketch also apply to the work of colleagues? *Is it universal/ variable/ Context-dependent?*
- Do you think it possible to describe an ML-workflow in general terms? *What are textbook phrases vs. real steps?*
- How well do you know the work of your colleagues/ partners? *Is there any documentation? Best practice for comparability?*
- How do you inform each other about your work? *Level of details, individual steps, particular parameter, issues and limitations*
- How difficult is it to become familiar with the work of other researchers in your domain?

Self-assessment

Experts reported their perception of their personal skills and obstacles in the context of ML.

- Opening question: Would you describe ML as one of your core areas?
- Do you feel you have a good overview of the development process?
- Would you consider yourself to be an ML expert? If not, who is an ML expert to you? How would you define that?
- Optionally, if not yet covered: Would you describe the development as black-box/opaque?

Closing questions

Explicit questions, e. g. concerning engineering challenges were placed at the very end to avoid priming. Further, participants had the opportunity to provide final comments and to pose questions.

- Opening Question: What do you think are the biggest challenges in ML? *Unused potentials? Limitations?*
- What would be a helpful support to improve the daily work?
- If you could wish for anything – what would that be?
- Do you have any further comments on any of the questions asked?
- Do you have questions about the interview, the evaluation or the requirements analysis in general?

7.2 Workflows

A simplified depiction of WF2 can already be found in the manuscript itself. The present section accordingly provides illustrations of WF1, WF3, WF4 and WF5, in Figure 4, Figure 2, Figure 3, and Figure 5, respectively.

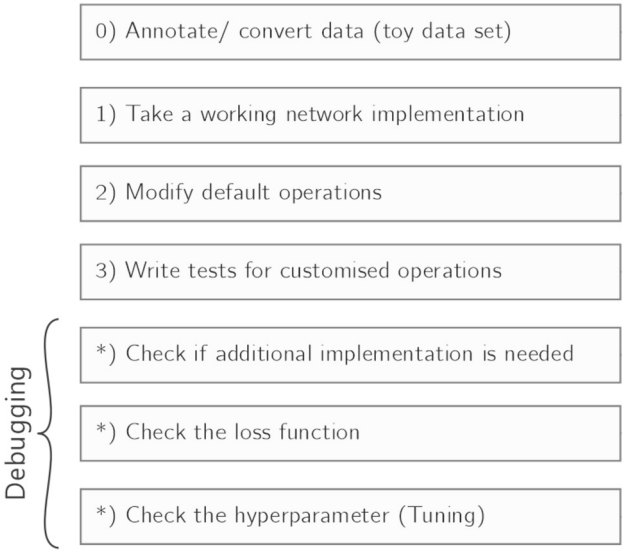

Figure 2: Workflow No. 3 (WF3) – Example ML Workflow sketched by one of the experts in the Think-Aloud task.

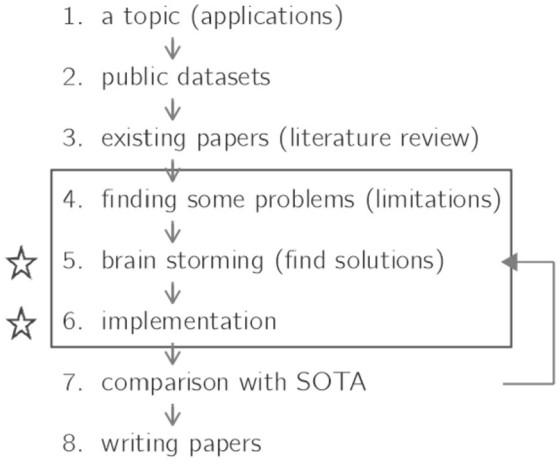

Figure 3: Workflow No. 4 (WF4) – Example ML Workflow sketched by one of the experts in the Think-Aloud task.

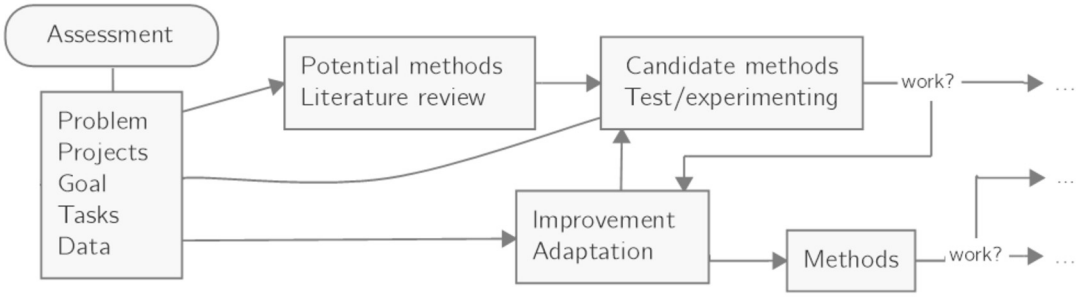

Figure 4: Workflow No. 1 (WF1) – Example ML Workflow sketched by one of the experts in the Think-Aloud task.

1. Data Collection & Processing
   - Task → Classification – similarity of classes
   - Feature selection
2. Apply Baseline
   - Simple model
   - Standard hyperparams

     → if performance on training set / loss is good

     → else try another baseline (simple one)

3. Improve: Regularization
   - Architecture:
     - Batch / Layer Norm.
     - Reg Loss / term

4. Data Augmentation

   → Make more data examples (if imbalance in classes)

   → Make training data more similar to test setting

Figure 5: Workflow No. 5 (WF5) – Example ML Workflow sketched by one of the experts in the Think-Aloud task.