

Operationalizing Machine Learning: An Interview Study

Shreya Shankar*, Rolando Garcia*, Joseph M. Hellerstein, Aditya G. Parameswaran

University of California, Berkeley

{shreyashankar,rogarcia,hellerstein,adityapp}@berkeley.edu

*Co-first authors

ABSTRACT

Organizations rely on machine learning engineers (MLEs) to operationalize ML, i.e., deploy and maintain ML pipelines in production. The process of operationalizing ML, or MLOps, consists of a continual loop of (i) data collection and labeling, (ii) experimentation to improve ML performance, (iii) evaluation throughout a multi-staged deployment process, and (iv) monitoring of performance drops in production. When considered together, these responsibilities seem staggering—how does anyone do MLOps, what are the unaddressed challenges, and what are the implications for tool builders?

We conducted semi-structured ethnographic interviews with 18 MLEs working across many applications, including chatbots, autonomous vehicles, and finance. Our interviews expose three variables that govern success for a production ML deployment: Velocity, Validation, and Versioning. We summarize common practices for successful ML experimentation, deployment, and sustaining production performance. Finally, we discuss interviewees’ pain points and anti-patterns, with implications for tool design.

1 INTRODUCTION

As Machine Learning (ML) models are increasingly incorporated into software, a nascent sub-field called *MLOps* (short for ML Operations) has emerged to organize the “set of practices that aim to deploy and maintain ML models in production reliably and efficiently” [4, 77]. It is widely agreed that MLOps is hard. Anecdotal reports claim that 90% of ML models don’t make it to production [76]; others claim that 85% of ML projects fail to deliver value [69].

At the same time, it is unclear *why* MLOps is hard. Our present-day understanding of MLOps is limited to a fragmented landscape of white papers, anecdotes, and thought pieces [14, 18, 20, 21, 34, 45], as well as a cottage industry of startups aiming to address MLOps issues [27]. Early work by Sculley et al. attributes MLOps challenges to “technical debt”, due to which there is “massive ongoing maintenance costs in real-world ML systems” [64]. Most successful ML deployments seem to involve a “team of engineers who spend a significant portion of their time on the less glamorous aspects of ML like maintaining and monitoring ML pipelines” [54]. Prior work has studied general practices of data analysis and science [30, 49, 62, 82], without considering MLOps challenges of productionizing models.

There is thus a pressing need to bring clarity to MLOps, specifically in identifying what MLOps typically involves—across organizations and ML applications. A richer understanding of best practices and challenges in MLOps can surface gaps in present-day processes and better inform the development of next-generation tools. Therefore, we conducted a semi-structured interview study of ML engineers (MLEs), each of whom has worked on ML models in production. We sourced 18 participants from different organizations and applications (Table 1) and asked them open-ended questions to understand their workflow and day-to-day challenges.

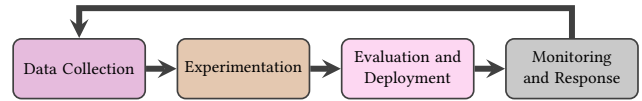


Figure 1: Routine tasks in the ML engineering workflow.

We find that MLEs perform **four routine tasks**, shown in Figure 1: (i) data collection, (ii) experimentation, (iii) evaluation and deployment, and (iv) monitoring and response. Across tasks, we observe **three variables** that dictate success for a production ML deployment: Velocity, Validation, and Versioning.¹ We describe common MLOps practices, grouped under overarching findings:

ML engineering is very experimental in nature (Section 4.3). As mentioned earlier, various articles claim that it is a problem for 90% of models to never make it to production [76], but we find that this statistic is misguided. The nature of constant experimentation is bound to create many versions, a small fraction of which (i.e. “the best of the best”) will make it to production. Thus it is beneficial to prototype ideas quickly, by making minimal changes to existing workflows, and demonstrate practical benefits early—so that bad models never make it far.

Operationalizing model evaluation requires an active organizational effort (Section 4.4). Popular model evaluation “best practices” do not do justice to the rigor with which organizations think about deployments: they generally focus on using one typically-static held-out dataset to evaluate the model on [38] and a single ML metric choice (e.g., precision, recall) [1, 2]. We find that MLEs invest significant resources in maintaining multiple up-to-date evaluation datasets and metrics over time—especially ensuring that data sub-populations of interest are adequately covered.

Non-ML rules and human-in-the-loop practices keep models reliable in production (Section 4.5). We find that MLEs prefer simple ideas, even if it means handling multiple versions: for example, rather than leverage advanced techniques to minimize distribution shift errors [15, 83], MLEs would simply create new models, retrained on fresh data. MLEs ensured that deployments were reliable via strategies such as on-call rotations, model rollbacks, or elaborate rule-based guardrails to avoid incorrect outputs.

In Section 5, we discuss recurring MLOps challenges across all tasks. We express these pain points as tensions and synergies between our three “V” variables—for example, undocumented “tribal knowledge” about pipelines (Section 5.2.4) demonstrates a tension between velocity (i.e., quickly changing the pipeline in response to a bug) and well-executed versioning (i.e., documenting every change). We conclude the description of each pain point with a discussion of opportunities for future tools.

¹Our Three Vs of MLOps aren’t meant to be confused with the Three Vs of Big Data (Volume, Variety, Velocity) [61]. The first authors learned of the Big Data Vs after drafting the MLOps Vs and were surprised to find similarities around volume/versioning and velocity.

	Role	Company Size	Application
p1	MLE Manager	Large	Autonomous vehicles
p2	MLE	Medium	Autonomous vehicles
p3	MLE	Small	Computer hardware
p4	MLE	Medium	Retail
p5	MLE Manager	Large	Ads
p6	MLE	Large	Cloud computing
p7	MLE	Small	Finance
p8	MLE	Small	NLP
p10	MLE	Small	OCR + NLP
p11	MLE Manager	Medium	Banking
p12	MLE	Large	Cloud computing
p13	MLE	Small	Bioinformatics
p14	MLE	Medium	Cybersecurity
p15	MLE	Medium	Fintech
p16	MLE	Small	Marketing and analytics
p17	MLE	Medium	Website builder
p18	MLE	Large	Recommender systems
p19	MLE Manager	Large	Ads

Table 1: Anonymized description of interviewees. Small companies have fewer than 100 employees; medium-sized companies have 100-1000 employees, and large companies have 1000 or more employees.

2 RELATED WORK

Several books and papers in the traditional software engineering literature describe the need for DevOps, a combination of software developers and operations teams, to streamline the process of delivering software in organizations [13, 37, 39, 40]. Similarly, *MLOps*, or DevOps principles applied to machine learning, has emerged from the rise of machine learning (ML) application development in software organizations. MLOps is a nascent field, where most existing papers give definitions and overviews of MLOps, as well as its relation to ML, software engineering, DevOps, and data engineering [28, 34, 44, 73]. MLOps poses unique challenges because of its focus on developing, deploying, and sustaining models, or artifacts that need to reflect data as data changes over time [59, 65, 67]. We discuss work related to MLOps workflows, challenges and interview studies for ML.

MLOps Workflow. The MLOps workflow involves supporting data collection and processing, experimentation, evaluation and deployment, and monitoring and response, as shown in Figure 1. Several research papers and companies have proposed tools to accomplish various tasks in the workflow, such as data pre-processing [22, 58, 60] and experiment tracking [6, 74, 81]. Crankshaw et al. studied the problem of model deployment and low-latency prediction serving [12]. With regards to validating changes in *production* systems, some researchers have studied CI (Continuous Integration) for ML and proposed preliminary solutions—for example, `ease.ml/ci` streamlines data management and proposes unit tests for overfitting [3], Garg et al. survey different MLOps tools [19], and some papers introduce tools to perform validation and monitoring in production ML pipelines [9, 31, 63].

MLOps Challenges. Sculley et al. were early proponents that production ML systems raise special challenges and can be hard to

maintain over time, based on their experience at Google [64]. Since then, several research projects have emerged to explore and tackle individual challenges in the MLOps workflow. For example, some discuss the need to manage data provenance and training context for model debugging purposes [8, 17, 24, 50]. Others describe the challenges of handling state and ensuring reproducibility (i.e., “managing messes”) while using computational notebooks [23, 42, 66]. Additionally, data distribution shifts have been technically but not operationally studied—i.e., how humans debug such shifts in practice [46, 51, 57, 71, 78]. Rather than focus on a single pain point, Lee et al. analyze challenges across ML workflows on an open-source ML platform [36]. Similarly, Xin et al. [79] analyze ML pipelines at Google to understand typical model configurations and retraining patterns. Polyzotis et al. [54, 55] survey challenges centric to data management for machine learning deployments. Paleyes et al. review published reports of individual ML deployments and survey common challenges [52]. Our study instead focuses on issues across the production workflow (i.e., MLOps practices and challenges) as opposed to individual pain-points, identified by interviewing those who are most affected by it—the ML engineers.

Data Science and ML-Related Interview Studies. Kandel et al. [30] interview data analysts at enterprises, focusing on broader organizational contexts like we do; however, MLOps workflows and challenges extend beyond data analysis. Other studies build on Kandel et al.’s work, exploring aspects such as collaboration, code practices, and tools [32, 33, 49, 53, 82], all centered on general data analysis and data science, as opposed to transitioning workflows in ML to production. Many ML-related interview studies focus on a single tool, task, or challenge in the workflow—for example, AutoML [75, 80], data iteration [25], model training [72], minimizing bias in ML models [26, 35, 43], and building infrastructure for ML pipelines [47]. Sambasivan et al. [62] study data quality issues during machine learning, as opposed to challenges in MLOps. Other ML-related interview studies focus on specific applications of ML, such as medicine [56], customer service [16], and interview processing [7]. Some interview studies report on software engineering practices for ML development; however, they focus only on a few applications and primarily on engineering, not operational, challenges [5, 41]. Our interview study aims to be both broad and focused: we consider many applications and companies, but is centered around the engineers that perform MLOps tasks, with an eye towards highlighting both engineering and operational practices and challenges. Additionally, our focus is on learning how models are deployed and sustained in production—we discover this by interviewing ML practitioners directly.

3 METHODS

Following review by our institution’s review board, we conducted an interview study of 18 ML Engineers (MLEs) working across a wide variety of sectors to learn more about their first-hand experiences serving and maintaining models in production.

3.1 Participant Recruitment

We recruited persons who were responsible for the development, regular retraining, monitoring and deployment of any ML model in

production. A description of the 18 MLEs (22% female-identifying²) is shown in Table 1. The MLEs we interviewed varied in their educational backgrounds, years of experience, roles, team size, and work sector. Recruitment was conducted in rounds over the course of an academic year (2021-2022). In each round, between three to five candidates were reached by email and invited to participate. We relied on our professional networks and open calls posted on MLOps channels in Discord³, Slack⁴, and Twitter to compile a roster of candidates. The roster was incrementally updated roughly after every round of interviews, integrating information gained from the concurrent coding and analysis of transcripts (Section 3.3). Recruitment rounds were repeated until we reached saturation on our findings [48].

3.2 Interview Protocol

With each participant, we conducted semi-structured interviews over video call lasting 45 to 75 minutes each. Over the course of the interview, we asked descriptive, structural, and contrast questions abiding by ethnographic interview guidelines [68]. The questions are listed in Appendix A. Specifically, our questions spanned six categories: (1) the type of ML task(s) they work on; (2) the approach(es) they use for developing and experimenting on models; (3) how and when they transition from development/experimentation to deployment; (4) how they evaluate their models prior to deployment; (5) how they monitor their deployed models; and (6) how they respond to issues or bugs that may emerge during deployment or otherwise. We covered these categories to get an overarching understanding of an ML lifecycle deployment, from conception to production sustenance.

Participants received a written consent form before the interview, and agreed to participate free of compensation. As per our agreement, we automatically transcribed the interviews using Zoom software. In the interest of privacy and confidentiality, we did not record audio or video of the interviews. Transcripts were redacted of personally identifiable information before being uploaded to a secured drive in the cloud. More information about the transcripts can be found in Appendix B.

3.3 Transcript Coding & Analysis

Following a grounded theory approach [10, 70], we employed open and axial coding to analyze our transcripts. We used MaxQDA, a common qualitative analysis software package for coding and comparative analysis. During a coding pass, two study personnel independently read interview transcripts closely to group passages into codes or categories. Coding passes were either top-down or bottom-up, meaning that codes were derived from theory or induced from interview passages, respectively. Between coding passes, study personnel met to discuss surprises and other findings, and following consensus, the code system was revised to reflect changes to the emerging theory. Coding passes were repeated until reaching convergence. More information about the codes is shown in

Appendix C, including a list of the most frequently occurring codes (Table 3) and co-occurring codes (Figure 4).

4 MLOPS PRACTICES: OUR FINDINGS

In this section, we present information about common practices in production ML deployments that we learned from the interviews. First, we describe common tasks in the production ML workflow in Section 4.1. Next, we introduce the Three Vs of MLOps, grounding both the discussion of findings and the challenges that we will explain in Section 5. Then in Section 4.3, we describe the strategies ML engineers leverage to produce successful experiment ideas. In Section 4.4, we discuss organizational efforts to effectively evaluate models. Finally, in Section 4.5, we investigate the hacks ML engineers use to sustain high performance in productions ML pipelines.

4.1 Tasks in the Production ML Lifecycle

We characterized ML engineers' workflows into four high-level tasks, each of which employ a wide variety of tools. We briefly describe each task in turn, and elaborate on them as they arise in our findings below.

Data Collection and Labeling. Data collection spans sourcing new data, wrangling data from sources into a centralized repository, and cleaning data. Data labeling can be outsourced (e.g., Mechanical Turk) or performed in-house with teams of annotators. Since descriptions and interview studies of data collection, analysis, wrangling and labeling activities can be found in related papers [11, 29, 30, 62], we focus our summary of findings on the other three tasks.

Feature Engineering and Model Experimentation. ML engineers typically focus on improving ML performance, measured via metrics such as accuracy or mean-squared-error. Experiments can be data-driven or model-driven; for example, an engineer can create a new feature or change the model architecture from tree-based to neural network-based.

Model Evaluation and Deployment. A model is typically evaluated by computing a metric (e.g., accuracy) over a collection of labeled data points hidden at training time, or a *validation dataset*, to see if its performance is better than what the currently-running production model achieved during its evaluation phase. Deployment involves reviewing the proposed change, possibly staging the change to increasing percentages of the population, or A/B testing on live users, and keeping records of any change to production in case of a necessary rollback.

ML Pipeline Monitoring and Response. Monitoring ML pipelines and responding to bugs involve tracking live metrics (via queries or dashboards), slicing and dicing sub-populations to investigate prediction quality, patching the model with non-ML heuristics for known failure modes, and finding in-the-wild failures and adding them to the evaluation set.

4.2 Three Vs of MLOps: Velocity, Validation, Versioning

When developing and pushing ML models to production, three properties of the workflow and infrastructure dictate how successful deployments will be: Velocity, Validation, and Versioning, discussed in turn.

²The skewed gender distribution is one possible indicator of sampling bias. We openly and actively recruited female-identifying MLEs to mitigate some sampling bias.

³mlops.discord.com

⁴mlops-community.slack.com

Velocity. Since ML is so experimental in nature, it's important to be able to prototype and iterate on ideas quickly (e.g., go from a new idea to a trained model in a day). ML engineers attributed their productivity to development environments that prioritized high experimentation velocity and debugging environments that allowed them to test hypotheses quickly (P1, P3, P6, P10, P11, P14, P18).

Validation. Since errors become more expensive to handle when users see them, it's good to test changes, prune bad ideas, and proactively monitor pipelines for bugs as early as possible (P1, P2, P5, P6, P7, P10, P14, P15, P18). P1 said: "The general theme, as we moved up in maturity, is: how do you do more of the validation earlier, so the iteration cycle is faster?"

Versioning. Since it's impossible to anticipate all bugs before they occur, it's helpful to store and manage multiple versions of production models and datasets for querying, debugging, and minimizing production pipeline downtime. ML engineers responded to buggy models in production by switching the model to a simpler, historical, or retrained version (P6, P8, P10, P14, P15, 18).

4.3 Machine Learning Engineering is Very Experimental, Even in Production

ML engineering, as a discipline, is highly experimental and iterative in nature, especially compared to typical software engineering. Contrary to popular negative sentiment around the large numbers of experiments and models that don't make it to production, we found that it's actually okay for experiments and models not to make it to production. What matters is making sure ideas can be prototyped and validated quickly—so that bad ones can be pruned away immediately. While there is no substitute for on-the-job experience to learn how to choose successful projects (P5), we document some self-reported strategies from our interviewees.

4.3.1 Good project ideas start with collaborators. Project ideas, such as new features, came from or were validated early by domain experts, data scientists and analysts who had already performed a lot of exploratory data analysis. P14 and P17 independently recounted successful project ideas that came from asynchronous conversations on Slack: P17 said, "I look for features from data scientists, [who have ideas of] things that are correlated with what I'm trying to predict." Solely relying on other collaborators wasn't enough, though—P5 mentioned that they "still need to be pretty proactive about what to search for."

Some organizations explicitly prioritized cross-team collaboration as part of their culture. P11 said:

We really think it's important to bridge that gap between what's often, you know, a [subject matter expert] in one room annotating and then handing things over the wire to a data scientist—a scene where you have no communication. So we make sure there's both data science and subject matter expertise representation [on our teams].

To foster a more collaborative culture, P16 discussed the concept of "building goodwill" with other teams through tedious tasks that weren't always explicitly a part of company plans:

Sometimes we'll fix something [here and] there to like build some good goodwill, so that we can call on them in the future...I do this stuff as I have to do it, not because I'm really passionate about doing it.

4.3.2 Iterate on the data, not necessarily the model. Several participants recommended focusing on experiments that provide additional context to the model, typically via new features (P5, P6, P11, P12, P14, P16, P17, P18, P19). P17 mentioned that most ML projects at their organization centered around adding new features. P14 mentioned that one of their current projects was to move feature engineering pipelines from Scala to SparkSQL (a language more familiar to ML engineers and data scientists), so experiment ideas could be coded and validated faster. P11 noted that iterating on the data, not the model, was preferable because it resulted in faster velocity:

I'm gonna start with a [fixed] model because it means faster [iterations]. And often, like most of the time empirically, it's gonna be something in our data that we can use to kind of push the boundary...obviously it's not like a dogmatic We Will Never Touch The Model, but it shouldn't be our first move.

Prior work has also identified the importance of data work [62].

4.3.3 Account for diminishing returns. At many organizations (especially larger companies), deployment can occur in stages—i.e., first validated offline, then deployed to 1% of production traffic, then validated again before a deployment to larger percentages of traffic. Some interviewees (P5, P6, P18) explained that experiment ideas typically have diminishing performance gains in later stages of deployment. As a result, P18 mentioned that they would initially try multiple ideas but focus only on ideas with the largest performance gains in the earliest stage of deployment; they emphasized the importance of "validat[ing] ideas offline...[to make] productivity higher." P19 corroborated this by saying end-to-end staged deployments could take several months, making it a high priority to kill ideas with minimal gain in early stages to avoid wasting future time. Additionally, to help with validating early, many engineers discussed the importance of a sandbox for stress-testing their ideas (P1, P5, P6, P11, P12, P13, P14, P15, P17, P18, P19). For some engineers, this was a single Jupyter notebook; others' organizations had separate sandbox environments to load production models and run ad-hoc queries.

4.3.4 Small changes are preferable to larger changes. In line with software best practices, interviewees discussed keeping their code changes as small as possible for multiple reasons, including faster code review, easier validation, and fewer merge conflicts (P2, P5, P6, P10, P11, P18, P19). Additionally, changes in large organizations were primarily made in config files instead of main application code (P1, P2, P5, P10, P12, P19). For example, instead of editing parameters directly in a Python model training script, it was preferable to edit a config file (e.g., JSON or YAML) of parameters instead and link the config file to the model training script.

P19 described how, as their team matured, they edited the model code less: "Eventually it was the [DAG] pipeline code which changed more...there was no reason to touch the [model] code...everything is config-based." P5 mentioned that several of their experiments

involved “[taking] an existing model, modify[ing] [the config] with some changes, and deploying it within an existing training cluster.” Supporting a config-driven development was important, P1 said, otherwise bugs might arise when promoting the experiment idea to production:

People might forget to, when they spawn multiple processes, to do data loading in parallel, they might forget to set different random seeds, especially [things] you have to do explicitly many times...you’re talking a lot about these small, small things you’re not going to be able to catch [at deployment time] and then of course you won’t have the expected performance in production.

Because ML experimentation requires many considerations to yield correct results—e.g., setting random seeds, accessing the same versions of code libraries and data—constraining engineers to config-only changes can reduce the number of bugs.

4.4 Operationalizing Model Evaluation is an Active Effort

We found that MLEs described intensive model evaluation efforts at their companies to keep up with data changes, product and business requirement changes, user changes, and organizational changes. The goal of model evaluation is to prevent repeated failures and bad models from making it to production while maintaining velocity—i.e., the ability for pipelines to quickly adapt to change.

4.4.1 Validation datasets should be dynamic. Many engineers reported processes to analyze live failure modes and update the validation datasets to prevent similar failures from happening again (P1, P2, P5, P6, P8, P11, P15, P16, P17, P18). P1 described this process as a departure from what they had learned in academia: “You have this classic issue where most researchers are evaluat[ing] against fixed data sets...[but] most industry methods change their datasets.” We found that these dynamic validation sets served two purposes: (1) the obvious goal of making sure the validation set reflects live data as much as possible, given new learnings about the problem and shifts in the aggregate data distribution, and (2) the more subtle goal of addressing localized shifts that subpopulations may experience (e.g., low accuracy for a specific label).

The challenge with (2) is that many subpopulations are typically unforeseen; many times they are discovered post-deployment. To enumerate them, P11 discussed how they systematically bucketed their data points based on the model’s error and created validation sets for each underperforming bucket:

Some [of the metrics in my tool] are standard, like a confusion matrix, but it’s not really effective because it doesn’t drill things down [into specific subpopulations that users care about]. Slices are user-defined, but sometimes it’s a little bit more automated. [During offline evaluation, we] find the error bucket that [we] want to drill down, and then [we] either improve the model in very systematic ways or improve [our] data in very systematic ways.

Rather than follow an anticipatory approach of constructing different failure modes in the offline validation phase—e.g., performance drops in subpopulations users might care deeply about—like

P11 did, P8 offered a reactive strategy of spawning a new dataset for each observed live failure: “Every [failed prediction] gets into the same queue, and 3 of us sit down once a week and go through the queue...then our [analysts] collect more [similar] data.” This new dataset was then used in the offline validation phase in future iterations of the production ML lifecycle.

While processes to dynamically update the validation datasets ranged from human-in-the-loop to frequent synthetic data construction (P6), we found that higher-stakes applications of ML (e.g., autonomous vehicles), created separate teams to manage the dynamic evaluation process. P1 said:

We had to move away from only aggregate metrics like MAP towards the ability to curate scenarios of interest, and then validate model performance on them specifically. So, as an example, you can’t hit pedestrians, right. You can’t hit cyclists. You need to work in roundabouts. You have a base layer of ML performance and the performance is not perfect...what you need to be able to do in a mature MLOps pipeline is go very quickly from user recorded bug, to not only are you going to fix it, but you also have to be able to drive improvements to the stack by changing your data based on those bugs.

Although the dynamic evaluation process might require many humans in the loop—a seemingly intense organizational effort—engineers thought it was crucial to have. When asked why they invested a lot of energy into their dynamic process, P11 said: “I guess it was always a design principle—the data is [always] changing.”

4.4.2 Validation systems should be standardized. The dynamic nature of validation processes makes it hard to effectively maintain versions of such processes, motivating efforts to standardize them. Several participants recalled instances of bugs stemming from inconsistent definitions of successful validation—i.e., where different engineers on their team evaluated models differently, causing unexpected changes to live performance metrics (P1, P3, P4, P5, P6, P7, P17). For instance, P4 lamented that every engineer working on a particular model had a cloned version of the main evaluation notebook, with a few changes. The inconsistent requirements for promoting a model to production caused headaches while monitoring and debugging, so their company instated a new effort to standardize evaluation scripts.

Although other MLOps practices highlighted the synergy between velocity and validating early (Section 4.3.3), we found that standardizing the validation system exposed a tension between velocity (i.e., being able to promote models quickly) and validating early, or eliminating the possibility of some bugs at deployment time. Since many validation systems needed to frequently change, as previously discussed, turnaround times for code review and merges to the main branch often could not keep up with the new tests and collections of data points added to the validation system. So, it was easier for engineers to fork and modify the evaluation system. However, P2 discussed that the decrease in velocity was worth it for their organization when they standardized evaluation:

We have guidelines on how to run eval[uation] comprehensively when any particular change is being made. Now there is a merge queue, and we have to make sure

that we process the merge queue in order, and that improvements are actually also reflected in subsequent models, so it requires some coordination. We'd much rather gate deploying a model than deploy a model that's bad. So we tend to be pretty conservative [now].

A standardized evaluation system also reduced friction in deploying ML in large companies and high-stakes settings. P5 discussed that for some models, deployments needed approvals across the organization, and it was much harder to justify a deployment with a custom or ad-hoc evaluation process: “At the end of the day, it’s all a business-driven decision...for something that has so much revenue riding on it, [you can’t have] a subjective opinion on whether [your] model is better.”

4.4.3 Spread a deployment across multiple stages, and evaluate at each stage. Several organizations, particularly those with many customers, had a multi-stage deployment process for new models or model changes, progressively evaluating at each stage (P3, P5, P6, P7, P8, P12, P15, P17, P18, P19). P6 described the staged deployment process as:

In [the large companies I’ve worked at], when we deploy code it goes through what’s called a staged deployment process, where we have designated test clusters, [stage 1] clusters, [stage 2] clusters, then the global deployment [to all users]. The idea here is you deploy increasingly along these clusters, so that you catch problems before they’ve met customers.

Each organization had different names for its stages (e.g., test, dev, canary, staging, shadow, A/B) and different numbers of stages in the deployment process (usually between one and four). The stages helped invalidate models that might perform poorly in full production, especially for brand-new or business-critical pipelines. P15 recounted an initial chatbot product launch using their staged deployment process, claiming it successfully “made it” because they were able to catch failures and update the model in early stages:

We spent a long time very slowly, ramping up the model to very small percentages of traffic and watching what happened. [When there was a failure mode,] a product person would ping us and say: hey, this was kind of weird, should we create a rule around this [suggested text] to filter this out?

Of particular note was one type of stage, the shadow stage—where predictions were generated live but not surfaced to users, that came before a deployment to a small fraction of live users. P14 described how they used the shadow stage to assess how impactful new features could be:

So if we’re testing out a new idea and want to see, what would the impact be for this new set of features without actually deploying that into production, we can deploy that in a type of shadow mode where it’s running alongside the production model and making predictions. We track all the metrics for [both] models in [our data lake]...so we can compare them easily.

Shadow mode had other use cases—for instance, P15 discussed how shadow mode was used to convince other stakeholders (e.g., product managers, business analysts) that a new model or change

to an existing model was worth putting in production. P19 mentioned that they use shadow mode to invalidate experiment ideas that would eventually fail. But shadow mode alone wasn’t a substitute for all stages of deployment—P6 said, “[in the early stage], we don’t have a good sample of how the model is going to behave in production”—requiring the multiple stages. Additionally, in products that have a feedback loop (e.g., recommender systems), it is impossible to evaluate the model in shadow mode because users do not interact with shadow predictions.

4.4.4 ML evaluation metrics should be tied to product metrics. Multiple participants stressed the importance of evaluating metrics critical to the product, such as click-through rate or user churn rate, rather than ML-specific metrics alone like MAP (P5, P7, P15, P16, P11, P17, P18, P19). The need to evaluate product-critical metrics stemmed from close collaboration with other stakeholders, such as product managers and business operators. P11 felt that a key reason many ML projects fail is that they don’t measure metrics that will yield the organization value:

Tying [model performance] to the business’s KPIs (key performance indicators) is really important. But it’s a process—you need to figure out what [the KPIs] are, and frankly I think that’s how people should be doing AI. It [shouldn’t be] like: hey, let’s do these experiments and get cool numbers and show off these nice precision-recall curves to our bosses and call it a day. It should be like: hey, let’s actually show the same business metrics that everyone else is held accountable to to our bosses at the end of the day.

Since product-specific metrics are, by definition, different for different ML models, it was important for engineers to treat choosing the metrics as an explicit step in their workflow and align with other stakeholders to make sure the right metrics were chosen. For example, P16 said that for every new ML project they work on, their “first task is to figure out, what are customers actually interested in, or what’s the metric that they care about.” P17 said that every model change in production is validated by the product team: “if we can get a statistically significant greater percentage [of] people to subscribe to [the product], then [we can fully deploy].”

For some organizations, a consequence of tightly coupling evaluation to product metrics was an additional emphasis on important customers during evaluation (P6, P10). P6 described how, at their company, experimental changes that increased aggregate metrics could sometimes be prevented from going to production:

There’s an [ML] system to allocate resources for [our product]. We have hard-coded rules for mission critical customers. Like at the beginning of Covid, there were hospital [customers] that we had to save [resources] for.

Participants who came from research or academia noted that tying evaluation to the product metrics was a different experience. P6 commented:

I think about where the business will benefit from what we’re building. We’re not just shipping off fake wins, like we’re really in the value business. You’ve got to see value from AI in your organization in order to feel like [our product] was worth it to you, and I guess that’s

a mindset that we really ought to have [as a broader community].

4.5 Sustaining Models Requires Deliberate Software Engineering and Organizational Practices

Here, we present a list of strategies ML engineers employed during monitoring and debugging phases to sustain model performance post-deployment.

4.5.1 Create new versions: frequently retrain on and label live data. Production ML bugs can be detected by tracking pipeline performance, measured by metrics like prediction accuracy, and triggering an alert if there is a drop in performance that exceeds some pre-defined threshold. On-call ML engineers noted that reacting to an ML-related bug in production often took a long time, motivating them to find alternative strategies to quickly restore performance (P1, P7, P8, P10, P14, P15, P17, P19). P14 mentioned automatically retraining the model every day so model performance would not suffer for longer than a day:

Why did we start training daily? As far as I'm aware, we wanted to start simple—we could just have a single batch job that processes new data and we wouldn't need to worry about separate retraining schedules. You don't really need to worry about if your model has gone stale if you're retraining it every day.

Retraining cadences ranged from hourly (P18) to every few months (P17) and were different for different models within the same organization (P1). None of the participants interviewed reported any scientific procedure for determining the cadence; the retraining cadences were set in a way that streamlined operations for the organization in the easiest way. For example, P18 mentioned that “retraining takes about 3 to 4 hours, so [they] matched the cadence with it such that as soon as [they] finished any one model, they kicked off the next training [job].”

Some engineers reported an inability to retrain unless they had freshly labeled data, motivating their organizations to set up a team to frequently label live data (P1, P8, P10, P11, P16). P10 reported that a group within their company periodically collected new documents for their language models to fine-tune on. P11 mentioned an in-house team of junior analysts to annotate the data; however, a problem was that these annotations frequently conflicted and the organization did not know how to reconcile the noise.

4.5.2 Maintain old versions as fallback models. Another way to minimize downtime when a model is known to be broken is to have a fallback model to revert to—either an old version or simple version. P19 said: “if the production model drops and the calibration model is still performing within a [specified] range, we'll fall back to the calibration model until someone will fix the production model.” P18 mentioned that it was important to keep some model up and running, even if they “switched to a less economic model and had to just cut the losses.”

4.5.3 Maintain layers of heuristics. P14 and P15 each discussed how their models are augmented with a final, rule-based layer to keep live predictions more stable. For example, P14 mentioned

working on an anomaly detection model and adding a heuristics layer on top to filter the set of anomalies that surface based on domain knowledge. P15 discussed one of their language models for a customer support chatbot:

The model might not have enough confidence in the suggested reply, so we don't return [the recommendation]. Also, language models can say all sorts of things you don't necessarily want it to—another reason that we don't show suggestions. For example, if somebody asks when the business is open, the model might try to quote a time when it thinks the business is open. [It might say] “9 am”, but the model doesn't know that. So if we detect time, then we filter that [reply]. We have a lot of filters.

Constructing such filters was an iterative process—P15 mentioned constantly stress-testing the model in a sandbox, as well as observing suggested replies in early stages of deployment, to come up with filter ideas. Creating filters was a more effective strategy than trying to retrain the model to say the right thing; the goal was to keep some version of a model working in production with little downtime. This combination of modern model-driven ML and old-fashioned rule-based AI indicates a need for managing filters (and versions of filters) in addition to managing learned models. The engineers we interviewed managed these artifacts themselves.

4.5.4 Validate data going in and out of pipelines. While participants reported that model parameters were typically “statically” validated before deploying to full production, features and predictions were continuously monitored for production models (P1, P2, P6, P8, P14, P16, P17, P18, P19). Several metrics were monitored—P2 discussed hard constraints for feature columns (e.g., bounds on values), P6 talked about monitoring completeness (i.e., fraction of non-null values) for features, P16 mentioned embedding their pipelines with “common sense checks,” implemented as hard constraints on columns, and P8 described schema checks—making sure each data item adheres to an expected set of columns and their types.

While rudimentary data checks were embedded in most systems, P6 discussed that it was hard to figure out what higher-order data checks to compute:

Monitoring is both metrics and then a predicate over those metrics that triggers alerts. That second piece doesn't exist—not because the infrastructure is hard, but because no one knows how to set those predicate values...for a lot of this stuff now, there's engineering headcount to support a team doing this stuff. This is people's jobs now; this constant, periodic evaluation of models.

Some participants discussed using black-box data monitoring services but lamented that their alerts did not prevent failures (P7, P14). P7 said:

We don't find those metrics are useful. I guess, what's the point in tracking these? Sometimes it's really to cover my ass. If someone [hypothetically] asked, how come the performance dropped from X to Y, I could go back in the data and say, there's a slight shift in the

user behavior that causes this. So I can do an analysis of trying to convince people what happened, but can I prevent [the problem] from happening? Probably not. Is that useful? Probably not.

While basic data validation was definitely useful for the participants, many of the participants expressed pain points with existing techniques and solutions, which we discuss further in Section 5.1.2.

4.5.5 Keep it Simple. Many participants expressed an aversion to complexity, preferring to rely on simple models and algorithms whenever possible (P1, P2, P6, P7, P11, P12, P14, P15, P16, P17, P19). P7 described the importance of relying on a simple training and hyperparameter search algorithm:

In finance, we always split data by time. The thing I [learned in finance] is, don't exactly try to tune the hyperparameters too much, because that just overfits to historic data.

P7 discussed choosing tree-based models over deep learning models for their ease of use, which simplified post-deployment maintenance: “I can probably do the same thing with neural nets. But it’s not worth it. [After] deployment it just doesn’t make any sense at all.” However, other participants chose to use deep learning as a means of simplifying their pipelines (P1, P16). For instance, P16 described training a small number of higher-capacity models rather than a separate model for each target: “There were hundreds of products that [customers] were interested in, so we found it easier to instead train 3 separate classifiers that all shared the same underlying embedding...from a neural network.”

While there was no universally agreed-upon answer to a question as broad as, “should I use deep learning?” we found a common theme in how participants leveraged deep learning models. Specifically, for ease of post-deployment maintenance (e.g., an ability to retroactively debug pipelines), outputs of deep learning models were typically human-interpretable (e.g., image segmentation, object recognition, probabilities or likelihoods as embeddings). P1 described a push at their company to rely *more* on neural networks:

A general trend is to try to move more into the neural network, and to combine models wherever possible so there are fewer bigger models. Then you don't have these intermediate dependencies that cause drift and performance regressions...you eliminate entire classes of bugs and and issues by consolidating all these different piecemeal stacks.

4.5.6 Organizationally Supporting ML Engineers Requires Deliberate Practices. Our interviewees reported various organizational processes for sustaining models as part of their ML infrastructure. P6, P12, P14, P16, P18, and P19 described *on-call processes* for supervising production ML models. For each model, at any point in time, some ML engineer would be on call, or primarily responsible for it. Any bug or incident observed (e.g., user complaint, pipeline failure) would receive a ticket, created by the on-call engineer, and be placed in a queue. On-call rotations lasted a week or two weeks. At the end of a shift, an engineer would create an incident report—possibly one for each bug—detailing major issues that occurred and how they were fixed.

Additionally, P6, P7, P8, P10, P12, P14, P15, P16, P18, and P19 mentioned having a *central queue of production ML bugs* that every engineer added tickets to and processed tickets from. Often this queue was larger than what engineers could process in a timely manner, so they assigned priorities to tickets. Finally, P6, P7, P10, and P15 discussed having *Service Level Objectives (SLOs)*, or commitments to minimum standards of performance, for pipelines in their organizations. For example, an pipeline to classify images could have an SLO of 95% accuracy. A benefit of using the SLO framework for ML pipelines is a clear indication of whether a pipeline is performing well or not—if the SLO is not met, the pipeline is broken, by definition.

5 MLOPS CHALLENGES AND OPPORTUNITIES

In this section, we enumerate common pain points and anti-patterns observed across interviews. We discuss each pain point as a tension or synergy between the Three Vs (Section 4.2). At the end of each pain point, we describe our takeaways of ideas for future tools and research. Finally, in Section 5.3, we characterize layers of the MLOps tool stack for those interested in building MLOps tools.

5.1 Pain Points in Production ML

We focus on four themes that we didn’t know before the interviews: the mismatch between development and production environments, handling a spectrum of data errors, the ad-hoc nature of ML bugs, and long validation processes.

5.1.1 Mismatch Between Development and Production Environments.

While it is important to create a separate development environment to validate ideas before promoting them to production, it is also necessary to minimize the discrepancies between the two environments. Otherwise, unanticipated bugs might arise in production (P1, P2, P6, P8, P10, P13, P14, P15, P18). **Creating similar development and production environments exposes a tension between velocity and validating early:** development cycles are more experimental and move faster than production cycles; however, if the development environment is significantly different from the production environment, it’s hard to validate (ideas) early.

We discuss three examples of point points caused by the environment mismatch—data leakage, Jupyter notebook philosophies, and code quality.

Data Leakage. A common issue was *data leakage*—i.e., assuming during training that there is access to data that does not exist at serving time—an error typically discovered after the model was deployed and several incorrect live predictions were made. Anticipating any possible form of data leakage is tedious and hinders velocity; thus, sometimes leakage was retroactively checked during code review (P1). The nature of data leakage ranged across reported bugs—for example, P18 recounted an instance where embedding models were trained on the same split of data as a downstream model, P2 described a *class imbalance* bug where they did not have enough labeled data for a subpopulation at training time (compared to its representation at serving time), and P15 described a bug in which *feedback delay* (time between making a live prediction and

getting its ground-truth label) was ignored while training. Different types of data leakage resulted in different magnitudes of ML performance drops: for example, in a pipeline with daily retraining, feedback delays could prevent retraining from succeeding because of a lack of new labels. However, in P18's embedding leakage example, the resulting model was slightly more overfitted than expected, yielding lower-than-expected performance in production but not completely breaking.

Strong Opinions on Jupyter Notebooks. Participants described strongly opinionated and different philosophies with respect to how to use Jupyter notebooks in their workflows. Jupyter notebooks were heavily used in development to support high velocity, which we did not find surprising. However, we were surprised that although participants generally acknowledged worse code quality in notebooks, some participants preferred to use them in production to minimize the differences between their development and production environments. P6 mentioned that they could debug quickly when locally downloading, executing, and manipulating data from a production notebook run. P18 remarked on the modularization benefits of a migration from a single codebase of scripts to notebooks:

We put each component of the pipeline in a notebook, which has made my life so much easier. Now [when debugging], I can run only one specific component if I want, not the entire pipeline... I don't need to focus on all those other components, and this has also helped with iteration.

On the other hand, some participants strongly disliked the idea of notebooks in production (P10, P15). P15 even went as far as to philosophically discourage the use of notebooks in the development environment: "Nobody uses notebooks. Instead, we all work in a shared code base, which is both the training and serving code base and people kick off jobs in the cloud to train models." Similarly, P10 recounted a shift at their company to move any work they wanted to reproduce or deploy out of notebooks:

There were all sorts of manual issues. Someone would, you know, run something with the wrong sort of inputs from the notebook, and I'm [debugging] for like a day and a half. Then [I'd] figure out this was all garbage. Eight months ago, we [realized] this was not working. We need[ed] to put in the engineering effort to create [non-notebook] pipelines.

The anecdotes on notebooks identified conflicts between competing priorities: (1) Notebooks support high velocity and therefore need to be in development environments, (2) Similar development and production environments prevents new bugs from being introduced, and (3) It's easy to make mistakes with notebooks in production, e.g., running with the wrong inputs; copy-pasting instead of reusing code. Each organization had different rankings of these priorities, ultimately indicating whether or not they used notebooks in production.

Non-standardized Code Quality. We found code quality and review practices to be non-standardized and inconsistent across development and production environments. Some participants described organization-wide production coding standards for specific languages (P2, P5), but even the most mature organizations did not

have ML-specific coding guidelines for experiments. Generally, experimental code (in development) was not reviewed, but changes to production went through a code review process (P1, P5). Participants felt that code review wasn't too useful, but they did it to adhere to software best practices (P1, P3, P5, P10). P5 mentioned that "it's just really not worth the effort; people might catch some minor errors". P10 hypothesized that the lack of utility came from difficulty of code review:

It's tricky. You use a little bit of judgment as to where things might go wrong, and you maybe spend more time sort of reviewing that. But bugs will go to production, and [as long as they're not] that catastrophic, [it's okay.]

Code review and other good software engineering practices might make deployments less error-prone. However, because ML is so experimental in nature, they can be significant barriers to velocity; thus, many model developers ignore these practices (P1, P6, P11). P6 said:

I used to see a lot of people complaining that model developers don't follow software engineering [practices]. At this point, I'm feeling more convinced that they don't follow software engineering [practices]—[not] because they're lazy, [but because software engineering practices are] contradictory to the agility of analysis and exploration.

Takeaway. We believe there's an opportunity to create virtualized infrastructure specific to ML needs with similar development and production environments. Each environment should build on the same foundation but supports different modes of iteration (i.e., high velocity in development). Such tooling should also track the discrepancies between environments and minimize the likelihood that discrepancy-related bugs arise.

5.1.2 Handling A Spectrum of Data Errors. As alluded to in Section 4.5.4, we found that ML engineers struggled to handle the spectrum of data errors: hard \rightarrow soft \rightarrow drift (P5, P6, P8, P11, P14, P16, P17, P18, P19). Hard errors are obvious and result in clearly "bad predictions", such as when mixing or swapping columns or when violating constraints (e.g., a negative age). Soft errors, such as a few null-valued features in a data point, are less pernicious and can still yield reasonable predictions, making them hard to catch and quantify. Drift errors occur when the live data is from a seemingly different distribution than the training set; these happen relatively slowly over time. One pain point mentioned by the interviewees was that different types of data errors require different responses, and it was not easy to determine the appropriate response. Another issue was that requiring practitioners to manually define constraints on data quality (e.g., lower and upper bounds on values) was not sustainable over time, as employees with this knowledge left the organization.

The most commonly discussed pain point was *false-positive alerts*, or alerts triggered even when the ML performance is adequate. Engineers often monitored and placed alerts on each feature or input column and prediction or output column (P5, P6, P8, P11, P14, P16, P17, P18, P19). Engineers automated schema checks and bounds to catch hard errors, and they tracked distance metrics (e.g., KL divergence) between historical and live features to catch soft

and drift errors. If the number of metrics tracked is so large, even with only a handful of columns, the probability that at least one column violates constraints is high!

Taking a step back, the purpose of assessing data quality *before* serving predictions is to validate early. **Correctly monitoring data quality demonstrates the conflict between validating early and versioning**—if data validation methods flag a broken data point, which in turn rejects the corresponding prediction made by the main ML model, some backup plan or fallback model version (Section 4.5) is necessary. Consequently, an excessive number of false-positive alerts leads to two pain points: (1) unnecessarily maintaining many model versions or simple heuristics, which can be hard to keep track of, and (2) a lower overall accuracy or ML metric, as baseline models might not serve high-quality predictions (P14, P19).

Dealing with Alert Fatigue. A surplus of false-positive alerts led to fatigue and silencing of alerts, which could miss actual performance drops. P8 said “people [were] getting bombed with these alerts.” P14 mentioned a current initiative at their company to reduce the alert fatigue:

Recently we’ve noticed that some of these alerts have been rather noisy and not necessarily reflective of events that we care about triaging and fixing. So we’ve recently taken a close look at those alerts and are trying to figure out, how can we more precisely specify that query such that it’s only highlighting the problematic events?

P18 shared a similar sentiment, that there was “nothing critical in most of the alerts.” The only time there was something critical was “way back when [they] had to actually wake up in the middle of the night to solve it...the only time [in years].” When we asked what they did about the noncritical alerts and how they acted on the alerts, P18 said:

You typically ignore most alerts...I guess on record I’d say 90% of them aren’t immediate. You just have to acknowledge them [internally], like just be aware that there is something happening.

The alert fatigue typically materialized when engineers were on-call, or responsible for ML pipelines during a 7 or 14-day shift. P19 recounted how on-call rotations were dreaded amongst their team, particularly for new team members, due to the high rate of false-positive alerts:

The pain point is dealing with that alert fatigue and the domain expertise necessary to know what to act on during on-call. New members freak out in the first [on-call], so [for every rotation,] we have two members. One member is a shadow, and they ask a lot of questions.

P19 also discussed an initiative at their company to reduce the alert fatigue, ironically with another model:

The [internal tool] looks at different metrics for what alerts were [acted on] during the on-call...[the internal tool] tries to reduce the noise level, alert. It says, hey, this alert has been populated this like 1,000 times and ignored 45% of time. [The on-call member] will acknowledge whether we need to [fix] the issue.

Creating Meaningful Data Alerts is Challenging. If schema checks and rudimentary column bounds didn’t flag all the errors, and distance metrics between historical and live feature values flagged too many false positive errors, how could engineers find a “Goldilocks” alert setting?⁵ We organized the data-related issues faced by engineers into a hierarchy, from most frequently occurring to least frequently occurring:

- **Feedback delays:** Many participants said that ground-truth labels for live predictions often arrived after a delay, which could vary unpredictably (e.g., human-in-the-loop or networking delays) and thus caused problems for knowing real-time performance or retraining regularly (P2, P7, P8, P15, P17, P18). P7 felt strongly about the negative impact of feedback delays on their ML pipelines:

I have no idea how well [models] actually perform on live data. We do log the the [feature and output] data, but feedback is always delayed by at least 2 weeks. Sometimes we might not have feedback...so when we realize maybe something went wrong, it could [have been] 2 weeks ago, and yeah, it’s just straight up—we don’t even care...nobody is solving the label lag problem. It doesn’t make sense to me that a monitoring tool is not addressing this, because [it’s] the number one problem.

P8 discussed how they spent 2-3 years developing a human-in-the-loop pipeline to manually label live data as frequently as possible: “you want to come up with the rate at which data is changing, and then assign people to manage this rate roughly”. On the other hand, P17 and P19 both talked about how, when they worked on recommender systems, they did not have to experience feedback delay issues. P17 said: “With recommendations, it’s pretty clear whether or not we got it right because we get pretty immediate feedback. We suggest something, and someone’s like go away or they click it.”

- **Unnatural data drift:** Often, in production pipelines, data was missing, incomplete, or corrupted, causing model performance to sharply degrade (P3, P6, P7, P10, P16, P17). Several participants cited Covid as an example, but there are other (better) everyday instances of unnatural data drift. P6 described a bug where users had inconsistent definitions of the same word, complicating the deployment of a service to a new user. P7 mentioned a bug where data from users in a certain geographic region arrived more sporadically than usual. P10 discussed a bug where the format of raw data was occasionally corrupted: “Tables didn’t always have headers in the same place, even though they were the same tables.”
- **Natural data drift:** Surprisingly, participants didn’t seem too worried about slower, expected natural data drift over time—they noted that frequent model retrains solved this problem (P6, P7, P8, P12, P15, P16, P17). As an anecdote, we asked P17 to give an example of a natural data drift problem their company faced, and they could not think of a good example. P14 also said they don’t have natural data drift problems:

⁵Goldilocks and the Three Bears is a popular Western fairy tale. Goldilocks, the main character, looks for things that are not too big or not too small, things that are “just right.”

The model gets retrained every day, so we don't have the scenario of like: Oh, our models got stale and we need to re-train it because it's starting to make mistakes because data has drifted...fortunately we've never had to deal with [such a] scenario. Sometimes there are bad [training] jobs, but we can always effectively roll back to a different [model].

However, a few engineers mentioned that natural data shift could cause some hand-curated features and data quality checks to corrupt (P3, P6, P8). P6 discussed a histogram used to make a feature (i.e., converting a real-valued feature to a categorical feature) for an ML model—as data changed over time, the bucket boundaries became useless, resulting in buggy predictions. P8 described how, in their NLP models, the vocabulary of frequently-occurring words changed over time, forcing them to update their preprocessor functions regularly. Our takeaway is that any function that summarizes data—be it cleaning tools, preprocessors, features, or models—needs to be refit regularly.

Takeaway. Unfortunately, none of the participants reported having solved the Goldilocks ML alert problem at their companies. What metrics can be reliably monitored in real-time, and what criteria should trigger alerts to maximize precision and recall when identifying model performance drops? How can these metrics and alerting criteria—functions of naturally-drifting data—automatically tune themselves over time? We envision this to be an opportunity for new data management tools.

5.1.3 Taming the Long Tail of ML Pipeline Bugs. In the interviews, we gathered the sentiment that ML debugging is different from debugging during standard software engineering, where one can write test cases to cover the space of potential bugs. But for ML, if one can't categorize bugs effectively because every bug feels unique, how will they prevent future similar failures? Moreover, it's important to fix pipeline bugs as soon as possible to minimize downtime, and **a long tail of possible ML pipeline bugs forces practitioners to have high debugging velocity.** “I just sort of poked around until, at some point, I figured [it] out,” P6 said, describing their ad-hoc approach to debugging. Other participants similarly mentioned that they debug without a systematic framework, which could take them a long time (P5, P8, P10, P18).

While some types of bugs were discussed by multiple participants, such as accidentally flipping labels in classification models (P1, P3, P6, P11) and forgetting to set random seeds (P1, P12, P13), the vast majority of bugs described to us in the interviews were seemingly bespoke and not shared among participants. For example, P8 forgot to drop special characters (e.g., apostrophes) for their language models. P6 found that the imputation value for missing features was once corrupted. P18 mentioned that a feature of unstructured data type (e.g., JSON) had half of the keys' values missing for a “long time.”

Unpredictable Bugs; Predictable Symptoms. Interestingly, these one-off bugs from the long tail showed similar symptoms of failure. For instance, a symptom of unnatural data drift issues (defined in Section 5.1.2) was a large discrepancy between offline validation accuracy and production accuracy immediately after deployment (P1, P6, P14, P18). The similarity in symptoms highlighted the similarity in methods for isolating bugs; they were almost always some

variant of slicing and dicing data for different groups of customers or data points (P2, P6, P11, P14, P17, P19). P14 discussed tracking bugs for different slices of data and only drilling down into their queue of bugs when they observed “systematic mistakes for a large number of customers.” P2 did something similar, although they hesitated to call it debugging:

You can sort of like, look for instances of a particular [underperforming slice] and [debug]—although I'd argue that [it isn't] debugging as much as it is sampling the world for more data...maybe it's not a bug, and [it's] just [that] the model has not seen enough examples of some slice.

Paranoia Caused by ML Debugging Trauma. After several iterations of chasing bespoke ML-related bugs in production, ML engineers that we interviewed developed a sense of paranoia while evaluating models offline, possibly as a coping mechanism (P1, P2, P6, P15, P17, P19). P1 recounted a bug that was “impossible to discover” after a deployment to production:

ML [bugs] don't get caught by tests or production systems and just silently cause errors [that manifest as] slight reductions in performance. This is why [you] need to be paranoid when you're writing ML code, and then be paranoid when you're coding. I remember one example of a beefy PR with a lot of new additions to data augmentation...but the ground truth data was flipped. If it hadn't been caught in code review, it [would've been] almost impossible to discover. I can think of no mechanism by which we would have found this besides someone just curiously reading the code. [In production], it would have only [slightly] hurt accuracy.

It's possible that many of the bespoke bugs could be ignored if they didn't actually affect model performance. Tying this concept to the data quality issue, maybe all engineers needed to know was when model performance was suffering. But they needed to know *precisely* when models were underperforming, an unsolved question as discussed in Section 5.1.2. When we asked P1, “how do you know when the model is not working as well as you expect?”—they gave the following answer:

Um, yeah, it's really hard. Basically there's no surefire strategy. The closest that I've seen is for people to integrate a very high degree of observability into every part of their pipeline. It starts with having really good raw data, observability, and visualization tools. The ability to query. I've noticed, you know, so much of this [ad-hoc bug exploration] is just—if you make the friction [to debug] lower, people will do it more. So as an organization, you need to make the friction very low for investigating what the data actually looks like, [such as] looking at specific examples.

Takeaway. Our takeaway is that there is a chicken-and-egg problem in making it easier to tackle the long tail of ML bugs. To group the tail into higher-order categories—i.e., to know what bugs to focus on and what to throw out—we need to know when models are precisely underperforming; then we can map performance drops

to bugs. However, to know when models are precisely underperforming, given feedback delays and other data quality assessment challenges as described in Section 5.1.2, we need to be able to identify all the bugs in a pipeline and reason how much each one could plausibly impact performance. Breaking this cycle could be a valuable contribution to the production ML community and help alleviate challenges that stem from the long tail of ML bugs.

5.1.4 Multi-Staged Deployments Seemingly Take Forever. Multiple participants complained that end-to-end experimentation—the conception of an idea to improve ML performance to validating the idea—took too long (P7, P14, P16, P17, P18, P19). **This reveals the synergies between velocity and validating early:** if ideas can be invalidated in earlier stages of deployment, then overall velocity is increased. But sometimes a stage of deployment would take a long time to observe meaningful results—for example, P19 mentioned that at their company, the timeline for trying a new feature idea took over three months:

I don't have the exact numbers; around 40 or 50% will make it to initial launch. And then, either because it doesn't pass the legal or privacy or some other complexity, we drop about 50% of [the launched experiments]. We have to drop a lot of ideas.

The uncertainty of whether projects will be successful stemmed from the unpredictable, real-world nature of the experiments (P18, P19). P19 said that some features don't make sense after a few months, given the nature of how user behaviors change, which would cause an initially good idea to never fully and finally deploy to production:

You have to look at so many different metrics, and even for very experienced folks doing this process like dozen times, sometimes it's hard to figure out especially when the user's behavior changes very steadily. There's no sudden change [in evaluation metrics] because of one launch; it just usage patterns that change.

P18 offered an anecdote where their company's key product metrics changed in the middle of one of their experiments, causing them to kill a experiment that appeared to be promising (the original metric was improving):

It was causing a huge gain on the product metrics; it was definitely a green signal. But as for the product, metrics keep on rotating based on the company's priorities, you know. Is it the revenue at this point? Is it the total number of, let's say, installs? Or clicks at this particular point of time? They keep on changing with company's roadmap...

Takeaway. While most participants were unable to share exact information about the length of the staged deployment process and specific anecdotes about experiments they needed to cancel for privacy reasons, we found it interesting how different organizations had different deployment evaluation practices yet similar pain around failed project ideas due to the highly iterative, experimental nature of ML. We believe there is an opportunity for tooling to streamline ML deployments in this multi-stage pattern, to minimize wasted work and help practitioners predict the end-to-end gains for their ideas.

5.2 Observed MLOps Anti-Patterns

Here we report a list of MLOps anti-patterns observed in our interviews, or common potentially-problematic behaviors in the ecosystem around ML experiments and deployments.

5.2.1 Industry-Classroom Mismatch. P1, P5, P7, P11, and P16 each discussed some ML-related bugs they encountered early in their career, right after leaving school, that they knew how to avoid only after on-the-job experience. “I learned a lot of data science in school, but none of it was quite like all these things you're [asking],” P7 told us at the end of their interview. P5 said they did a lot of “learning by doing.” P11 provided further insight:

It was [hard to be], like, thrown into the wild, and have to learn all of this on the job. Coming out of [university in the US with a strong CS program], these are not things that anyone has ever taught right, at least in school...my mindset has always been a little bit more, I guess, practically oriented, even since the academic days, and that's not to say we had great mental models—for frameworks or playbooks—for doing this.

Our interviews with participants didn't focus on what specific skills they could have learned in the classroom that would have prepared them better for their jobs. We leave this to future study and collaborations between academia and industry.

5.2.2 Keeping GPUs Warm. P5 first mentioned the phrase “keeping GPUs warm”—i.e., running as many experiments as possible given computational resources, or making sure all GPUs were utilized at any given point in time:

One thing that I've noticed is, especially when you have as many resources as [large companies] do, that there's a compulsive need to leverage all the resources that you have. And just, you know, get all the experiments out there. Come up with a bunch of ideas; run a bunch of stuff. I actually think that's bad. You can be overly concerned with keeping your GPUs warm, [so much] so that you don't actually think deeply about what the highest value experiment is.

I think you can end up saving a lot more time—and obviously GPU cycles, but mostly end-to-end completion time—if you spend more efforts choosing the right experiment to run instead of [spreading yourself] thin. All these different experiments have their own frontier to explore, and all these frontiers have different options. I basically will only do the most important thing from each project's frontier at a given time, and I found that the net throughput for myself has been much higher.

In executing experiment ideas, we noticed a tradeoff between a guided search and random search. Random searches were more suited to parallelization—e.g., hyperparameter searches or ideas that didn't depend on each other. Although computing infrastructure could support many different experiments in parallel, the cognitive load of managing such experiments was too cumbersome for participants (P5, P10, P18, P19). In other words, **having high velocity means drowning in a sea of versions** of experiments. Rather, participants noted more success when pipelining learnings from one experiment into the next, like a guided search to find the

best idea (P5, P10, P18). P18 described their ideological shift from random search to guided search:

Previously, I tried to do a lot of parallelization. I used to take, like, 5 ideas and try to run experimentation in parallel, and that definitely not only took my time, but I also focused less. If I focus on one idea, a week at a time, then it boosts my productivity a lot more.

By following a guided search, engineers are, essentially, significantly pruning a tree of experiment ideas without executing them. Contrary to what they were taught in academia, P1 observed that some hyperparameter searches could be pruned early because hyperparameters had such little impact on the end-to-end pipeline:

I remember one example where the ML team spent all this time making better models, and it was not helping [overall performance]. Then everyone was so frustrated when one person on the controls team just tweaked one parameter [for the non-ML part of the pipeline], and [the end-to-end pipeline] worked so much better. Like we've invested all this infrastructure for hyperparameter tuning a experiment, and I'm like what is this. Why did this happen?

Our takeaway is that while it may seem like there are unlimited computational resources, developer time and energy is the limiting reagent for ML experiments. At the end of the day, experiments are human-validated and deployed. Mature ML engineers know their personal tradeoff between parallelizing disjoint experiment ideas and pipelining ideas that build on top of each other, ultimately yielding successful deployments.

5.2.3 Retrofitting an Explanation. Right from the first interview, participants discussed uncovering good results from experiments, productionizing changes, and then trying to reason why these changes worked so well (P1, P2, P7, P12). P1 said:

A lot of ML is is like: people will claim to have like principled stances on why they did something and why it works. I think you can have intuitions that are useful and reasonable for why things should be good, but the most defining characteristic of [my most productive colleague] is that he has the highest pace of experimentation out of anyone. He's always running experiments, always trying everything. I think this is relatively common—people just try everything and then backfit some nice-sounding explanation for why it works.

We wondered, why was it even necessary to have an explanation for why something worked? Why not simply accept that, unlike in software, we may not have elegant, principled reasons for successful ML experiments? P2 hypothesized that such retrofitted explanations could guide future experiment ideas over a longer horizon. Alternatively, P7 mentioned that their customers sometimes demanded explanations for certain predictions:

Do I know why? No idea. I have to convince people that, okay, we try our best. We try to [compute] correlations. We try to [compute] similarities. Why is it different? I have to make conjectures.

We realized that although they could be false, retrofitted explanations can help with collaboration and business goals. If they

satisfy customers and help organize teams around a roadmap of experiment ideas, maybe they are not so bad.

5.2.4 Undocumented Tribal Knowledge. P6, P10, P13, P14, P16, P17, and P19 each discussed pain points related to undocumented knowledge about ML experiments and pipelines amongst collaborators with more experience related to specific pipelines. Across interviews, it seemed like **high velocity created many versions, which made it hard to maintain up-to-date documentation**. P10 mentioned that there were parts of a pipeline that no one touched because it was already running in production, and the principal developer who knew most about it had left the company. P16 said that “most of the, like, actual models were trained before [their] time.” P14 described a “pipeline jungle” that was difficult to maintain:

You end up with this pipeline jungle where everything's super entangled, and it's really hard to make changes, because just to make one single change, you have to hold so much context in your brain. You're trying to think about like, okay this one change is gonna affect this system which affects this [other] system, [which creates]...the pipeline got to the point where it was very difficult to make even simple changes.

While writing down institutional knowledge can be straightforward to do once, P6 discussed that in the ML setting, they learn faster than they can document; moreover, people don't want to read so many different versions of documentation:

There are people in the team, myself included, that have been on it for several years now, and so there's some institutional knowledge embodied on the team that sometimes gets written down. But you know, even when it does get written down, maybe you will read them, but then, they kind of disappear to the ether.

Finally, P17 realized that poorly documented pipelines forced them to treat pipelines as black boxes: “Some of our models are pretty old and not well documented, so I don't have great expectations for what they should be doing.” Without intuition for how pipelines should perform, practitioner productivity can be stunted.

Takeaway. The MLOps anti-patterns described in this section reveal that ML engineering, as a field, is changing faster than educational resources can keep up. We see this as opportunities for new resources, such as classroom material (e.g., textbooks, courses) to prescribe the right engineering practices and rigor for the highly experimental discipline that is production ML, and automated documentation assistance for ML pipelines in organizations.

5.3 Characterizing the “MLOps Stack” for Tool Builders

MLOps tool builders may be interested in an organization of the dozens of tools, libraries, and services MLEs use to run ML and data processing pipelines. Although multiple MLEs reported having to “glue” open-source solutions together and having to build “homegrown” infrastructure as part of their work (P1, P2, P5, P6, P10, P12), an analysis of the various deployments reveals that tools

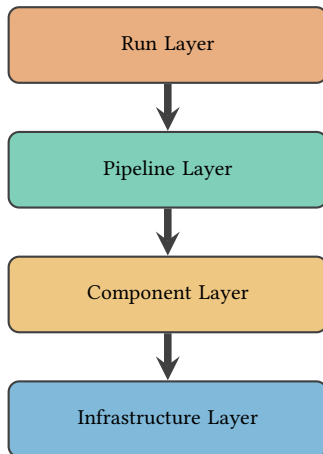


Figure 2: Layers of tools in the MLOps stack.

can be grouped into a stack of four layers, depicted in Figure 2 and discussed further in Appendix D. We discuss the four layers in turn:

- (1) **Run Layer:** A *run* is a record of an execution of an ML or data pipeline (and its components). Run-level data is often managed by data catalogs, model registries, and training dashboards.

Example Tools: Weights & Biases, MLFlow, Hive metastores, AWS Glue

- (2) **Pipeline Layer:** Finer-grained than a run, a pipeline further specifies the dependencies between artifacts and details of the corresponding computations. Pipelines can run ad-hoc or on a schedule. Pipelines change less frequently than runs, but more frequently than components.

Example Tools: Papermill, DBT, Airflow, TensorFlow Extended, Sagemaker

- (3) **Component Layer:** A component is an individual node of computation in a pipeline, often a script inside a managed environment. Some MLEs reported having an organization-wide “library of common components” for pipelines to use, such as feature generation and model training (P2, P6).

Example Tools: Python, Spark, PyTorch, TensorFlow

- (4) **Infrastructure Layer:** MLEs described a wide range of solutions, but most used cloud storage (e.g., S3), and GPU-backed cloud computing (AWS and GCP). Infrastructure changed far less frequently than other layers in the stack, but each change was more laborious and prone to wide-ranging consequences.

Example Tools: Docker, AWS, GCP

We found that MLEs used layers of abstraction (e.g., “config-based development”) as a way to manage complexity: most changes (especially high-velocity ones) were minor and limited to the Run Layer, such as selecting hyperparameters. As the stack gets deeper, changes become less frequent: MLEs ran training jobs daily but modified Dockerfiles occasionally. In the past, as MLOps tool builders, we (the authors) have incorrectly assumed uniform user access patterns across all layers of the MLOps stack. Tool builders may

want to pay attention to the layer(s) they are addressing and make sure they are not designing tools for the wrong layer(s).

Additionally, we noticed a high-level pattern in how interviewees discussed the tools they used: engineers seemed to prefer tools that significantly improved their experience with respect to the Three Vs (Section 4.2). For example, experiment tracking tools increased engineers’ speed of iterating on feature or modeling ideas (P14, P15)—a velocity virtue. In another example, feature stores (i.e., tables of derived columns for ML models) helped engineers debug models because they could access the relevant historical versions of features used in training such models (P3, P6, P14, P17)—a versioning virtue. MLOps tool builders may want to prioritize “10x” better experiences across velocity, validating early, or versioning for their products.

6 CONCLUSION

In this paper, we presented results from a semi-structured interview study of ML engineers spanning different organizations and applications to understand their workflow, best practices, and challenges. We found that successful MLOps practices center around having high velocity, validating as early as possible, and maintaining multiple versions of models for minimal production downtime. We reported on the experimental nature of production ML, aspects of effective model evaluation, and tips to sustain model performance over time. Finally, we discussed MLOps pain points and anti-patterns discovered in our interviews to inspire new MLOps tooling and research ideas.

ACKNOWLEDGEMENTS

We thank the interviewees for their valuable time and thoughtful responses. We are also grateful to Sarah Catanzaro for connecting us to some of the interviewees, and Alex Tamkin and Preetum Nakkiran for helpful suggestions. We acknowledge support from grants IIS-2129008, IIS-1940759, IIS-1940757, and CNS-1730628 awarded by the National Science Foundation, DOE Grant No. DE-SC0016934, an NDSEG Graduate Fellowship, an NSF Graduate Research Fellowship, funds from the Alfred P. Sloan Foundation, as well as EPIC lab sponsors: Adobe, Microsoft, Google, and Sigma Computing. Work was done while Hellerstein was on leave at Sutter Hill Ventures.

REFERENCES

- [1] Evaluating a model - advice for applying machine learning.
- [2] Single number evaluation metric - ml strategy.
- [3] Lionel Aguilar, David Dao, Shaoduo Gan, Nezihe Merve Gurel, Nora Hollenstein, Jiawei Jiang, Bojan Karlas, Thomas Lemmin, Tian Li, Yang Li, Susie Rao, Johannes Rausch, Cedric Renggli, Luka Rimanic, Maurice Weber, Shuai Zhang, Zhikuan Zhao, Kevin Schawinski, Wentao Wu, and Ce Zhang. Ease.ml: A lifecycle management system for mldev and ml ops. In *Conference on Innovative Data Systems Research (CIDR 2021)*, January 2021.
- [4] Sridhar Alla and Suman Kalyan Adari. What is ml ops? In *Beginning MLOps with MLFlow*, pages 79–124. Springer, 2021.
- [5] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300, 2019.
- [6] Lukas Biewald. Tracking with weights and biases www.wandb.com/, 2020.
- [7] Catherine Billington, Gonzalo Rivero, Andrew Jannett, and Jiating Chen. A machine learning model helps process interviewer comments in computer-assisted personal interview instruments: A case study. *Field Methods*, page 1525822X221107053, 2022.
- [8] Mike Brachmann, Carlos Bautista, Sonia Castelo, Su Feng, Juliana Freire, Boris Glavic, Oliver Kennedy, Heiko Müller, Rémi Rampin, William Spoth, and Ying

- Yang. Data debugging and exploration with vizier. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD '19*, page 1877–1880, New York, NY, USA, 2019. Association for Computing Machinery.
- [9] Eric Breck, Marty Zinkevich, Neoklis Polyzotis, Steven Whang, and Sudip Roy. Data validation for machine learning. In *Proceedings of SysML*, 2019.
- [10] Ji Young Cho and Eun-Hee Lee. Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences. *Qualitative report*, 19(32), 2014.
- [11] Xu Chu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang. Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD '16*, page 2201–2206, New York, NY, USA, 2016. Association for Computing Machinery.
- [12] Daniel Crankshaw, Xin Wang, Giulio Zhou, Michael J Franklin, Joseph E Gonzalez, and Ion Stoica. Clipper: A {Low-Latency} online prediction serving system. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, pages 613–627, 2017.
- [13] Christof Ebert, Gorka Gallardo, Josune Hernantes, and Nicolas Serrano. Devops. *Ieee Software*, 33(3):94–100, 2016.
- [14] Mihail Eric. Mlops is a mess but that's to be expected.
- [15] Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking importance weighting for deep learning under distribution shift. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11996–12007. Curran Associates, Inc., 2020.
- [16] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. What makes users trust a chatbot for customer service? an exploratory interview study. In *International conference on internet science*, pages 194–208. Springer, 2018.
- [17] Rolando Garcia, Eric Liu, Vikram Sreekanti, Bobby Yan, Anusha Dandamudi, Joseph E. Gonzalez, Joseph M Hellerstein, and Koushik Sen. Hindsight logging for model training. In *VLDB*, 2021.
- [18] Rolando Garcia, Vikram Sreekanti, Neeraja Yadwadkar, Daniel Crankshaw, Joseph E Gonzalez, and Joseph M Hellerstein. Context: The missing piece in the machine learning lifecycle. In *CFM*, 2018.
- [19] Satvik Garg, Pradyumn Pundir, Geetanjali Rathee, P.K. Gupta, Somya Garg, and Saransh Ahlawat. On continuous integration / continuous delivery for automated deployment of machine learning models using mlops. In *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 25–28, 2021.
- [20] Inc Gartner. Understanding mlops to operationalize machine learning projects.
- [21] Samadrita Ghosh. Mlops challenges and how to face them, Aug 2021.
- [22] Stefan Grafberger, Shubha Guha, Julia Stoyanovich, and Sebastian Schelter. Mlinspect: A data distribution debugger for machine learning pipelines. In *SIGMOD '21*, 2021.
- [23] Andrew Head, Fred Hohman, Titus Barik, Steven M. Drucker, and Robert DeLine. Managing messes in computational notebooks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.
- [24] Joseph M Hellerstein, Vikram Sreekanti, Joseph E Gonzalez, James Dalton, Akon Dey, Sreyashi Nag, Krishna Ramachandran, Sudhanshu Arora, Arka Bhat-tacharyya, Shirshanka Das, et al. Ground: A data context service. In *CIDR*, 2017.
- [25] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. Understanding and visualizing data iteration in machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020.
- [26] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, page 1–16, New York, NY, USA, 2019. Association for Computing Machinery.
- [27] Chip Huyen. Machine learning tools landscape v2 (+84 new tools), Dec 2020.
- [28] Meenu Mary John, Helena Holmström Olsson, and Jan Bosch. Towards mlops: A framework and maturity model. In *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 1–8. IEEE, 2021.
- [29] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3363–3372, 2011.
- [30] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. Enter-prise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2917–2926, 2012.
- [31] Daniel Kang, Deepti Raghavan, Peter Bailis, and Matei Zaharia. Model assertions for debugging machine learning.
- [32] Mary Beth Kery, Amber Horvath, and Brad A Myers. Variolite: Supporting exploratory programming by data scientists. In *CHI*, volume 10, pages 3025453–3025626, 2017.
- [33] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. Data scientists in software teams: State of the art and challenges. *IEEE Transactions on Software Engineering*, 44(11):1024–1038, 2017.
- [34] Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. Machine learning operations (mlops): Overview, definition, and architecture, 2022.
- [35] Po-Ming Law, Sana Malik, Fan Du, and Moumita Sinha. Designing tools for semi-automated detection of machine learning biases: An interview study. *arXiv preprint arXiv:2003.07680*, 2020.
- [36] Angela Lee, Doris Xin, Doris Lee, and Aditya Parameswaran. Demystifying a dark art: Understanding real-world machine learning model development, 2020.
- [37] Leonardo Leite, Carla Rocha, Fabio Kon, Dejan Milojicic, and Paulo Meirelles. A survey of devops concepts and challenges. *ACM Computing Surveys (CSUR)*, 52(6):1–35, 2019.
- [38] Zhiqiu Lin, Jia Shi, Deepak Pathak, and Deva Ramanan. The CLEAR benchmark: Continual LEArning on real-world imagery. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- [39] Mike Loukides. *What is DevOps?* " O'Reilly Media, Inc.", 2012.
- [40] Lucy Ellen Lwakatara, Terhi Kilamo, Teemu Karvonen, Tanja Sauvola, Ville Heikkilä, Juha Itkonen, Pasi Kuvaja, Tommi Mikkonen, Markku Oivo, and Casper Lassenius. Devops in practice: A multiple case study of five companies. *Information and Software Technology*, 114:217–230, 2019.
- [41] Lucy Ellen Lwakatara, Aiswarya Raj, J. Bosch, Helena Holmström Olsson, and Ivica Crnkovic. A taxonomy of software engineering challenges for machine learning systems: An empirical investigation. In *XP*, 2019.
- [42] Stephen Macke, Hongpu Gong, Doris Jung-Lin Lee, Andrew Head, Doris Xin, and Aditya Parameswaran. Fine-grained lineage for safer notebook interactions. *Proc. VLDB Endow.*, 14(6):1093–1101, sep 2021.
- [43] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.
- [44] Sasu Mäkinen, Henrik Skogström, Eero Laaksonen, and Tommi Mikkonen. Who needs mlops: What data scientists seek to accomplish and how can mlops help? In *2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN)*, pages 109–112. IEEE, 2021.
- [45] MLReef. Global mlops and ml tools landscape: Mlreef, Feb 2021.
- [46] Jose G. Moreno-Torres, Troy Raeder, Rocio Alaiiz-Rodriguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.
- [47] Dennis Muiruri, Lucy Ellen Lwakatara, Jukka K Nurminen, and Tommi Mikkonen. Practices and infrastructures for ml systems—an interview study in finnish organizations. 2022.
- [48] Michael Muller. Curiosity, creativity, and surprise as analytic tools: Grounded theory method. In *Ways of Knowing in HCI*, pages 25–48. Springer, 2014.
- [49] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019.
- [50] Mohamad Hossein Namaki, Avriella Floratou, Fotis Psallidas, Subru Krishnan, Ashvin Agrawal, Yinghui Wu, Yiwen Zhu, and Markus Weimer. Vamsa: Automated provenance tracking in data science scripts. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1542–1551, 2020.
- [51] Yaniv Ovadia, Emily Fertig, J. Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019.
- [52] Andrei Paleyes, Raoul-Gabriel Urma, and Neil D. Lawrence. Challenges in deploying machine learning: A survey of case studies. *ACM Comput. Surv.*, apr 2022. Just Accepted.
- [53] Samir Passi and Steven J Jackson. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–28, 2018.
- [54] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. Data management challenges in production machine learning. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1723–1726, 2017.
- [55] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. Data Lifecycle Challenges in Production Machine Learning: A Survey. *SIGMOD Record*, 47(2):12, 2018.
- [56] Luisa Pumplun, Mariska Fecho, Nihal Wahl, Felix Peters, Peter Buxmann, et al. Adoption of machine learning systems for medical diagnostics in clinics: qualitative interview study. *Journal of Medical Internet Research*, 23(10):e29301, 2021.
- [57] Stephan Rabanser, Stephan Günemann, and Zachary Chase Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *NeurIPS*, 2019.
- [58] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data*

- Bases, volume 11, page 269. NIH Public Access, 2017.
- [59] Cedric Renggli, Luka Rimanic, Nezihe Merve Gürel, Bojan Karlaš, Wentao Wu, and Ce Zhang. A data quality-driven view of mlops. *arXiv preprint arXiv:2102.07750*, 2021.
 - [60] E. Rezig et al. Dagger: A data (not code) debugger. In *CIDR*, 2020.
 - [61] Philip Russum et al. Big data analytics. *TDWI best practices report, fourth quarter*, 19(4):1–34, 2011.
 - [62] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. “everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.
 - [63] Sebastian Schelter et al. Automating large-scale data quality verification. In *PVLDB’18*, 2018.
 - [64] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *NIPS*, 2015.
 - [65] Shreya Shankar, Bernease Herman, and Aditya G. Parameswaran. Rethinking streaming machine learning evaluation. *ArXiv*, abs/2205.11473, 2022.
 - [66] Shreya Shankar, Stephen Macke, Sarah Chasins, Andrew Head, and Aditya Parameswaran. Bolt-on, compact, and rapid program slicing for notebooks. *Proc. VLDB Endow.*, sep 2023.
 - [67] Shreya Shankar and Aditya G. Parameswaran. Towards observability for production machine learning pipelines. *ArXiv*, abs/2108.13557, 2022.
 - [68] James P Spradley. *The ethnographic interview*. Waveland Press, 2016.
 - [69] Steve Nunez. Why ai investments fail to deliver, 2022. [Online; accessed 15-September-2022].
 - [70] Anselm Strauss and Juliet Corbin. Grounded theory methodology: An overview. 1994.
 - [71] Masashi Sugiyama et al. Covariate shift adaptation by importance weighted cross validation. In *JMLR*, 2007.
 - [72] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S. Tan. Ensemblematrix: Interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’09, page 1283–1292, New York, NY, USA, 2009. Association for Computing Machinery.
 - [73] Damian A Tamburri. Sustainable mlops: Trends and challenges. In *2020 22nd international symposium on symbolic and numeric algorithms for scientific computing (SYNASC)*, pages 17–23. IEEE, 2020.
 - [74] Manasi Vartak. Modeldb: a system for machine learning model management. In *HILDA ’16*, 2016.
 - [75] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. Human-ai collaboration in data science: Exploring data scientists’ perceptions of automated ai. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.
 - [76] Joyce Weiner. Why ai/data science projects fail: how to avoid project pitfalls. *Synthesis Lectures on Computation and Analytics*, 1(1):i–77, 2020.
 - [77] Wikipedia contributors. Mlops – Wikipedia, the free encyclopedia, 2022. [Online; accessed 15-September-2022].
 - [78] Olivia Wiles, Sven Goyal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. *ArXiv*, abs/2110.11328, 2021.
 - [79] Doris Xin, Hui Miao, Aditya Parameswaran, and Neoklis Polyzotis. Production machine learning pipelines: Empirical analysis and optimization opportunities. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2639–2652, 2021.
 - [80] Doris Xin, Eva Yiwei Wu, Doris Jung-Lin Lee, Niloufar Salehi, and Aditya Parameswaran. Whither automl? understanding the role of automation in machine learning workflows. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA, 2021. Association for Computing Machinery.
 - [81] M. Zaharia et al. Accelerating the machine learning lifecycle with mlflow. *IEEE Data Eng. Bull.*, 41:39–45, 2018.
 - [82] Amy X Zhang, Michael Muller, and Dakuo Wang. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23, 2020.
 - [83] Marvin Zhang, Henrik Marklund, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group shift. *CoRR*, abs/2007.02931, 2020.

A SEMI-STRUCTURED INTERVIEW QUESTIONS

In the beginning of each interview, we explained the purpose of the interview—to better understand processes within the organization for validating changes made to production ML models, ideally through stories of ML deployments. We then kickstarted the information-gathering process with a question to build rapport with the interviewee, such as *tell us about a memorable previous ML model deployment*. This question helped us isolate an ML pipeline or product to discuss. We then asked a series of open-ended questions:

- (1) **Nature of ML task**
 - What is the ML task you are trying to solve?
 - Is it a classification or regression task?
 - Are the class representations balanced?
- (2) **Modeling and experimentation ideas**
 - How do you come up with experiment ideas?
 - What models do you use?
 - How do you know if an experiment idea is good?
 - What fraction of your experiment ideas are good?
- (3) **Transition from development to production**
 - What processes do you follow for promoting a model from the development phase to production?
 - How many pull requests do you make or review?
 - What do you look for in code reviews?
 - What automated tests run at this time?
- (4) **Validation datasets**
 - How did you come up with the dataset to evaluate the model on?
 - Do the validation datasets ever change?
 - Does every engineer working on this ML task use the same validation datasets?
- (5) **Monitoring**
 - Do you track the performance of your model?
 - If so, when and how do you refresh the metrics?
 - What information do you log?
 - Do you record provenance?
 - How do you learn of an ML-related bug?
- (6) **Response**
 - What historical records (e.g., training code, training set) do you inspect in the debugging process?
 - What organizational processes do you have for responding to ML-related bugs?
 - Do you make tickets (e.g., Jira) for these bugs?
 - How do you react to these bugs?
 - When do you decide to retrain the model?

B INTERVIEW TRANSCRIPTS

Histograms of the number of codes and sentences in the interview transcripts are shown in Figures 3a and 3b, respectively.

C CODES

Across the interview transcripts, we had a total of 1766 coded segments, with exactly 600 unique codes. We organized codes into hierarchies. Table 3 shows the most frequently occurring codes, ordered by the number of distinct interviews the codes appeared

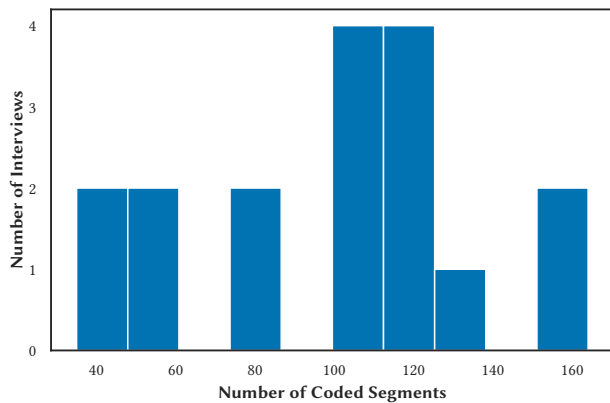
in (not the raw number of occurrences across all documents). Figure 4 displays the top five correlated codes for each top-level or parent code. Two codes are correlated if they occur within twenty sentences of each other.

D MLOPS TOOL STACK

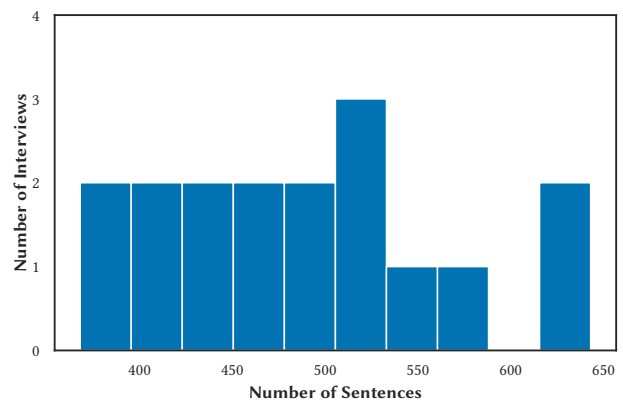
Table 2 shows common tools used by MLEs across layers of the stack and tasks in the production ML lifecycle.

	Data Collection	Experimentation	Evaluation and Deployment	Monitoring and Response
Run	Know what data is available and where it lives	Prototype ideas and track results	Catch errors in training (e.g., overfitting)	Track ML metrics over time
	Data catalogs, Amundsen, AWS Glue, Hive metastores	Weights & Biases, MLFlow, train/test set parameter configs, A/B test tracking tools		Dashboards, SQL, metric functions and window sizes
Pipeline	Regularly scheduled, possibly outsourced	Ad-hoc or user-triggered, hyperparameter search	Scheduled refresh of hold-out validation sets	Scheduled computation of metrics and triggered alerts
	In-house or outsourced annotators	AutoML	Github Actions, Travis CI, Prediction serving tools, Kafka, Flink	Prometheus, AWS Cloud-Watch
	Airflow, Kubeflow, Argo, Tensorflow Extended (TFX), Vertex AI, DBT			
Component	Sourcing, labeling, cleaning	Feature generation and selection, model training	Running model on hold-out validation set, model compression or rewrite, model serialization	Data validation, ML metric computation, tracing predictions
	Data cleaning tools	Tensorflow, MLlib, PyTorch, Scikit-learn, XG-Boost	C++, ONNX, OctoML, TVM, jobjlib, pickle	Scikit-learn metric functions, Great Expectations, Deequ
	Python, Pandas, Spark, SQL			
Infrastructure	Velocity	Velocity	Validate early	Versioning
	Annotation schema, cleaning criteria configs	Jupyter notebook setups, GPU	Edge devices, CPUs	Logging and observability services (e.g., Data-Dog)
	Cloud (e.g., AWS, GCP), compute clusters, storage (e.g., AWS S3, Snowflake), Docker, Kubernetes			

Table 2: Primary goals and tools for each layer in the MLOps stack and routine task in the ML engineering workflow.



(a) Histogram of number of coded segments in each interview.



(b) Histogram of number of sentences in each interview.

Figure 3: Interview transcript statistics. Each histogram has 10 equally-spaced buckets.

	Parent Code	Code	# Coded Segments	# Transcripts
1	known challenges	data drift/shift/skew	15	10
2	monitoring and response (+)	live monitoring	21	9
3	Python (+)	Jupyter	16	8
4	evaluation and deployment	build the infrastructure	16	8
5	data pipeline	data iteration, fresh data	15	8
6	fast & simple	high iteration speed, agile, rapid cycles	24	7
7	production bugs	debugging and bugs	18	7
8	tests	AB Testing	14	7
9	software development	pull request	13	7
10	data pipeline	pipeline on a schedule	13	7
11	operations	model training & retraining	12	7
12	data ingest	automated featurization	9	7
13	evaluation and deployment	CI/CD	8	7
14	trends	per-customer model and many customers	15	6
15	known challenges	feedback delay	13	6
16	evaluation and deployment	metrics and validation	12	6
17	metrics and validation (+)	accuracy	11	6
18	models	deep learning	9	6
19	sandboxing	offline demonstration of value	7	6
20	apps & use-cases	ranking	7	6

Table 3: Top 20 codes, ordered by the number of distinct transcripts the codes were mentioned in, descending.

