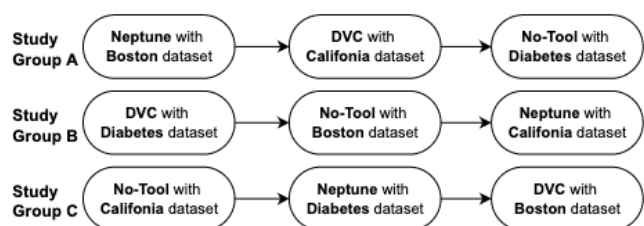


### Control Experiment Study Protocol

1. Protocol Title	ML Experiment Management Tools: Controlled Experiment
2. Background	An empirical study protocol for controlled experiments with ML experiment management tools
3. Primary Objectives	Empirical findings for use as a guide and requirement for designing new and improved ML experiment management tools.
4. Secondary Objectives	Investigate the effect of ML experiment management tools on user performance and provide insight into user preferences on tool paradigms.
5. Hypothesis	<p><b>Since we have observed that</b> specialized experiment management tools' effects on user performance are unknown <b>by</b> the support features they provide.</p> <p><b>We want to</b> establish the effect of these tools on user performance <b>when</b> using the tools vs. the baseline of the manual approach.</p> <p><b>Which should lead to</b> noticeably increased performance for the specialized tools vs. lower performance for the manual approach, <b>and the effect will be measured by</b> the ability of users to correctly and completely answer fact-based questions on assets managed by the subject tools..</p>
6. Variables	<p>Dependent &amp; independent &amp; extraneous</p> <p><b>Independent variables:</b></p> <ul style="list-style-type: none"><li>-The subject tools under consideration (Neptune. AI, DVC, and No-Tool)</li><li>-The ML dataset used in the experiment's ML tasks (Boston, California, and Diabetes datasets)</li></ul> <p><b>Dependent variables:</b></p> <ul style="list-style-type: none"><li>-The error rate: indicates how many wrong answers are provided in a given task for each tool.</li><li>- The completion rate: indicates how many questions were answered either wrongly or correctly for each tool.</li></ul> <p><b>Extraneous:</b></p> <ul style="list-style-type: none"><li>- We identify variables such as "implementation speed" as variables</li></ul>

	<p>that could affect our experiment. We mitigate the impact by providing a skeleton script that includes required code lines such as import statements.</p>
<p>7. Experiment Design</p>	<p>The primary phenomenon under investigation is the “effect of ML experiment management tools on users”.</p> <p>To evaluate this effect, we adopted the following steps:</p> <ul style="list-style-type: none"> <li>• Select the test subjects: We selected two tools representing different tool paradigms (Tool A and B). As a baseline for comparison, we consider No-Tool setup as a third option (Tool C).</li> <li>• We designed a series of tasks based on the ML workflow. Specifically, a supervised learning ML task to generate an ML model.</li> <li>• We designed 3 different tasks based on 3 different datasets (Dataset A, B, and C)</li> <li>• The experiment setup requires the participants to perform the tasks using different combinations of the 3 tools and datasets.</li> <li>• We split the participants into 3 groups, Group A, B and C.</li> <li>• We applied a <b>cross-over</b> design to effectively vary the tools and datasets for each of the study group as shown below</li> </ul>  <pre> graph LR     subgraph Study_Group_A [Study Group A]         A1([Neptune with Boston dataset]) --&gt; A2([DVC with California dataset])         A2 --&gt; A3([No-Tool with Diabetes dataset])     end     subgraph Study_Group_B [Study Group B]         B1([DVC with Diabetes dataset]) --&gt; B2([No-Tool with Boston dataset])         B2 --&gt; B3([Neptune with California dataset])     end     subgraph Study_Group_C [Study Group C]         C1([No-Tool with California dataset]) --&gt; C2([Neptune with Diabetes dataset])         C2 --&gt; C3([DVC with Boston dataset])     end </pre> <ul style="list-style-type: none"> <li>• During the experiment tasks, participants are required to answer fact-based questions to access how well they can track and retrieve assets with respective tools.</li> <li>• We used the responses to determine the dependent variables (i.e., error and completion rates).</li> </ul>

<p>8. Participant allocation &amp; Treatments</p>	<p>We recruit 15 participants based on the following selection criteria</p> <ul style="list-style-type: none"> <li>• Familiarity with ML frameworks</li> <li>• No prior experience with the ML tools.</li> </ul> <p>To improve the validity, we adopted a <b>cross-over</b> design and randomly divided our participants into three different study groups with 5 participants per group. For example, participants in group A received treatments in the order of 1, 2, and 3. In contrast, participants in study group B received treatments in the order of 2, 3, and 1 (See illustration under Experiment Design).</p> <p>For the random division, we assigned integer numbers 1 - 15 to the participants, randomized the numbers, and split the first 5, following, and last 5 numbers into separate groups.</p> <p>The study groups performed the experiment tasks with the same set of tools and datasets, but we varied the order of datasets/tools to avoid learning effects.</p>
<p>9. Experiment Materials</p>	<p>The experiment materials include</p> <p><b>Tutorials:</b> These are materials sent out to participants on how to set up the tools and quickly get started.</p> <p><b>Python scripts &amp; Config files:</b> To limit the effect of extraneous variables, we provide config files and initial working scripts that include code blocks required for the tasks.</p> <p><b>Instruction files:</b> These are materials guiding participants through the required steps. The instruction files are designed for each of the study groups.</p> <p>In the online appendix, we provide the materials above in the following directories:</p> <ul style="list-style-type: none"> <li>• Neptune-materials: This directory includes a short Neptune tutorial, the initial python scripts for the experiment phases using Neptune.ai to track ML assets, and instruction files for each of the participant group.</li> </ul>

	<ul style="list-style-type: none"> <li>• <b>DVC-materials:</b> This directory includes a short DVC tutorial, the initial python scripts and DVC config files for the experiment phases that used DVC to track and manage ML assets, and instruction files for each of the participant group</li> <li>• <b>No-Tool-materials:</b> This directory includes the initial python scripts for the experiment phases, which participants carried out without any supporting tool, and the instruction files for each of the participant group</li> </ul> <p><b>Questionnaire &amp; Responses:</b> We used Google Form to deliver the experiment information and questions to participants. We provide the following in the online appendix.</p> <ul style="list-style-type: none"> <li>• <b>Questionnaire:</b> Content of the Google Form used as a questionnaire for the controlled experiment. Some information, such as links and participants' emails, have been redacted to preserve anonymity. The questionnaire focused on two aspects: <ul style="list-style-type: none"> <li>◦ The fact-based questions to evaluate the user performance for each subject tool.</li> <li>◦ Questions to elicit users' opinions on tools paradigms and user preferences.</li> </ul> </li> <li>• <b>Response/data:</b> A collection of data from the controlled experiment as responses to our questionnaire.</li> </ul> <p><b>Analysis:</b> We provide the Excel spreadsheet used for the analysis of obtained responses.</p> <ul style="list-style-type: none"> <li>• <b>Statistical Data Analysis:</b> Raw data, extracted data, and statistical test analysis of quantitative data points.</li> </ul>
10. Participants Demographics	Total of 15 undergraduate students

	<table> <tr> <th colspan="2">Machine Learning Experience</th></tr> <tr> <td>&lt;6 months</td><td>40 %</td></tr> <tr> <td>6 - 12 months</td><td>60 %</td></tr> <tr> <th colspan="2">Experience with:</th></tr> <tr> <td>Git</td><td>100 %</td></tr> <tr> <td>ML Experiment Management Tools</td><td>0 %</td></tr> </table>	Machine Learning Experience		<6 months	40 %	6 - 12 months	60 %	Experience with:		Git	100 %	ML Experiment Management Tools	0 %
Machine Learning Experience													
<6 months	40 %												
6 - 12 months	60 %												
Experience with:													
Git	100 %												
ML Experiment Management Tools	0 %												
11. Data collection and Analysis	<p>We collected the participants' responses from Google Forms as a CSV file (provided in the appendix). The responses contain the following responses</p> <ul style="list-style-type: none"> <li>• Responses to factual questions in Table IV of the paper.</li> <li>• Responses to perception and paradigm-based questions in Table V</li> </ul> <p>For the responses to factual questions, we analyzed them by i) comparing the expected result values to the obtained responses to determine the error rates and ii) the number of attempted questions to determine the completion rates. The analysis material is provided in the "<b>Statistical data analysis_on_user_performance.xlsx</b>" file.</p> <p>For the responses to user perception questions, we adopt both qualitative (open-ended questions) and quantitative analysis (Likert-scale and multiple-choice questions). For the qualitative analysis of the open-ended questions, we applied thematic analysis. Specifically, we identified recurring and essential themes in the participants' responses and organized these themes (a.k.a. codes) in a hierarchy. For the quantitative analysis, we created descriptive statistics of corresponding responses.</p>												
12. Manipulation check.	<p>Since the validity of our conclusion is based on how accurately the participants' error and correctness rate reflect the underlying tools' support. We strongly advised and encouraged our participants not to cheat (e.g., repeating the experiment steps).</p>												
13. Statistical tests	<p>We performed statistical tests on our "performance-impact" result. We used Kruskal-Wallis, a non-parametric test for multi-group comparisons, suited for smaller</p>												

	<p>groups that are likely not normally distributed. We applied a Bonferroni correction to the significance threshold of 5% for three comparisons leading to a corrected threshold of 1.67%.</p>
<p>14. Construct &amp; Conclusion Validity</p>	<p>To ensure the experiment instructions and associated questions correctly reflect our intention to capture the usage of tools on user performance, we performed a series of dry runs to validate our experiment instruments. Specifically, we performed three different dry runs providing opportunities to improve the construct validity of our materials. We provide tutorial materials on the subject tools for participants 24 hours before the tests to ensure a correct understanding of used terminologies and questions. For the conclusion validity, we employed additional empirical methods (practitioner survey and user-opinion survey) to draw the conclusions from our study. Our experiment's findings align with those obtained from complementary surveys. Similarly, we employed appropriate statistical tests for the tools' performance comparison to mitigate validity threats due to smaller participant groups. In general, we argue that the methodology adopted was appropriate to obtain reliable insights regarding the effect of specialized tools on user performance.</p>