

# Interview Script:

## Welcome

Thank you for participating in our study of machine learning metadata search and usability. I'm going to go ahead and start the recording of our session.

[start recording]

Thank you for participating in this interview process and completing the survey questions and consent form.

[tech check]

Can you unmute and speak up to confirm the audio is working?

[internet check]

Can you make sure you have an internet browser tab open and ready to search before we begin?

[re-explain consent]

Now, I know you have already filled out your consent forms, I appreciate that, so I just want to confirm that you are ok with me using your recorded audio and video to review reactions and choices made when answering questions. We will only use the collected data for this project, and all data we gather will be stored on my computer and accessible to only myself and my PI. Is that still ok?

[check for verbal/non-verbal affirmation]

## Setup

[explain the setup]

For this technology setup, I will ask you to share your screen at one point during the interview, which is the "share" button at the bottom of your screen, so I can monitor your progress through certain tasks. After I confirm we are recording and you can share here, we should be ready to explain the process and then get started.

[check that they can share screen and Zoom is still recording]

I will be enabling closed captioning in Zoom for transcription purposes.

[check that closed captioning has been activated and is hidden]

You have kindly volunteered to participate in a study examining metadata and user preferences for machine learning geospatial datasets—specifically searchability and usability. We want to inquire about the interactions users have with machine learning metadata. This study is unfunded but will inform the Machine Learning Commons Croissant working group on machine learning dataset search behavior so they can translate the themes from this study to development of their metadata schema.

Again, we need to emphasize that we are not testing you but instead trying to understand how machine learning practitioners work with metadata. The interview questions are broadly written to cover any processes, resources, or issues related to the spatial machine learning dataset domain. Also, we can stop at any time for any reason, we want to make sure you are comfortable.

[explain testing process]

For this study I will be asking you to answer a series of questions that will require you to recall actions taken during your machine learning work, and also preferences you have regarding search. Additionally, I may ask you to perform several follow up tasks in which you look something up on your internet browser or show me on your screen. For each question I ask, please take your time in answering the question—there is no time limit. It helps to keep the datasets you work with in mind while answering all the questions about search and retrieval of machine learning data.

[check in before beginning]

Do you have any questions about what I've outlined or any other related questions?

[answer questions, if applicable]

## Interview questions

[participants will answer each question in linear order to be consistent]

### General questions

- Can you describe the area of machine learning that you work in?
  - How long have you been working in this area?
- Can you describe the type of training data you work with? (such as image, text, tabular data, etc.)

- The datasets you use for training, are they made from scratch or repurposed from other sources?
  - Do you create metadata for the data you process/receive?

### **Search for data**

- How often are you searching for data for machine learning purposes within your work?
- What tools and platforms **are you aware of** to search for geospatial machine learning datasets? (such as hugging face)
  - Moderator note: have them list off ones they are aware of
  - What tools/platforms **do you use** to search for geospatial machine learning datasets?
  - How do you search on that (confirmed) platform? Specifically, do you use tags for facets like languages, size, format, and/or do you filter by trending or popular?
  - Can you walk me through the last search you can recall for a geospatial machine learning dataset? (please go ahead and share your screen and talk through the steps of your search, starting with )

### **Interoperability**

- Can you describe any challenges you've had in using data from different dataset platforms (e.g., HuggingFace and Kaggle)?

### **Trust in data**

- How do you determine what is relevant to your research (using information like the metadata, title, contents, data card?)
  - How do you eventually pick a dataset for machine learning (is it quality, usability, number of likes, number of downloads, word of mouth?)
- Do you read the data cards of datasets?
  - (Moderator note: if confusion over what a data card is: data cards are “structured summaries of essential facts about various aspects of ML datasets needed by stakeholders across a dataset’s lifecycle for responsible AI development, and describe the content (What information to present), design (How to present information), and evaluation (Assess the efficacy of information)”)
    - What attributes do you look for within the data cards?
    - What attributes would you want to see that may be missing from metadata?

### **What is missing from search?**

- How would you improve dataset search for yourself?
  - What additions to tools such as google dataset search would benefit you when searching for machine learning datasets?

### **AI assistance in searching**

- Do you use large language model applications (e.g., ChatGPT) to assist you when searching for datasets?

- If so, can you describe how you search for datasets with them and then trust the returned results?

## Task-based questions

1. For the “Can you walk me through the last search you can recall for a geospatial machine learning dataset?” task:
  - i. Think Aloud Protocol from participant to elicit these questions:
    1. Why did you click on that dataset? What features were you interacting with? What filters did you have, if any?

## Post-test questions

1. On a scale of 1-5, 1 being very hard for you, 5 being very easy for you, how would you rate your overall ability to locate a relevant geospatial dataset?
  - a. Can you explain your rating?
2. Is there any part of the geospatial ML workflow pipeline that causes you issues or is not satisfactory?
  - a. Can you explain your answer?
3. Do you have any additional feedback or questions you would like to pose to the developers of geospatial metadata standards?

[conclude]

Alright, that concludes our questions and interview. Thank you for your participation, the information we collected will help us make recommendations to Croissant metadata standard developers and others in the machine learning metadata domain. Please let me know if you have any remaining questions. There will be no additional followup needed. Enjoy the rest of your day.

[end recording, interview over]