Cloud Computing Amazon SageMaker GroundTruth

Farid Afzali, Ph.D., P.Eng.

AWS SageMaker GroundTruth

- AWS SageMaker GroundTruth is delineated as a service for data labeling, pivotal in the preparation of datasets for machine learning.
- The term 'ground truth' is defined within the machine learning lexicon as the 'gold standard'. This standard is used as the benchmark for evaluating the predictive accuracy of machine learning models.
- Ground Truth serves to identify the correct classification or outcome that a machine learning model is expected to predict, which is essential for supervised learning where the model's output is compared against known labeled data.
- The service encompasses various data labeling tasks, such as:
  - Image Classification: Categorizing images into predefined classes.
  - Bounding Box: Identifying and marking the location of objects within images using rectangular shapes.
  - Semantic Segmentation: Labeling specific pixels of an image to differentiate between objects and background, thus understanding the image at the pixel level.
  - Label Verification: Ensuring the accuracy of labels provided to the data, which is a quality control step to confirm whether the assigned labels are correct or incorrect.

AWS SageMaker GroundTruth Tasks

- https://aws.amazon.com/sagemaker/data-labeling/features/
- **Bounding Boxes**: This task involves drawing rectangles over objects within images. The coordinates of the box edges are used to train models to locate and identify objects in other images. It's a common approach used in computer vision for object detection and localization.
- **Image Classification**: In this task, images are categorized into predefined labels or classes. Each image is assigned a label that represents the content of the image. This is used to train models to recognize and classify images into these learned categories.
- **Semantic Segmentation**: This involves labeling each pixel in an image with a class that denotes what object the pixel belongs to. Unlike bounding boxes that locate objects, semantic segmentation differentiates between objects at the pixel level, enabling detailed understanding and analysis of images. It's particularly useful in applications like autonomous driving where the precise location of objects is crucial.
- **Text Classification**: This task is about assigning predefined labels to text data. For instance, classifying email messages into 'spam' or 'not spam'. This form of classification helps in training models to understand and categorize text data based on content.
- **Custom Tasks**: SageMaker GroundTruth also allows for the creation of custom labeling tasks. This flexibility means that if a project has specific requirements not covered by the standard tasks, a custom workflow can be designed. For example, labeling audio data, video frames, or complex document analysis that may require a combination of labeling techniques. Custom tasks can be tailored with specific instructions and interfaces to meet the unique needs of any machine learning project.

Tasks Available Data Labeling Workforces

Public Mechanical Turks: Amazon SageMaker connects customers with a large, on-demand workforce known as Mechanical Turk workers. This global pool consists of approximately 500,000 contractors who are available 24/7 to perform data labeling tasks.

Private: Customers have the option to use a private team of labelers for their data labeling needs. This can include the customer's own private labelers, and all management and coordination are conducted through SageMaker GroundTruth. It is noted that there is no requirement for these labelers to have an IAM (AWS Identity and Access Management) or an Amazon account.

Vendors: SageMaker GroundTruth also offers access to professional data labeling services through a network of third-party vendors. These vendors are curated and can provide specialized data labeling services. This option is for customers seeking professional and potentially more sophisticated data labeling services that may not be available from the public Mechanical Turk workforce or private labelers.

Four Primary Categories of Data Sources

Image/Video:

- This category includes visual data, both still and moving. Such data is rich in content and can be used for a variety of applications including facial recognition, object detection, video analytics, and more. Analyzing image and video data often requires advanced techniques like computer vision and deep learning.

Text (Corpus):

- Textual data sources, or corpora, consist of collections of written texts, which could be in the form of books, articles, social media posts, etc. Text data is used for natural language processing (NLP) tasks such as sentiment analysis, topic modeling, and language translation.

Audio/Sound:

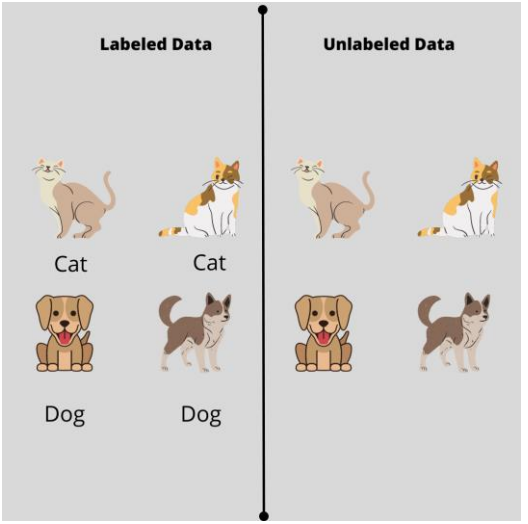
- Audio data sources encompass sounds, speech, and other auditory signals. Analyzing audio data can involve speech recognition, music classification, and environmental sound analysis, among other tasks. Techniques such as signal processing and deep learning are commonly employed here.

Time Series/Signals:

- This refers to data that is collected over time, often at regular intervals. Examples include stock prices, weather data, and physiological signals like heart rate. Time series analysis is crucial for forecasting, anomaly detection, and understanding temporal dynamics.

Labeled and Unlabeled Datasets

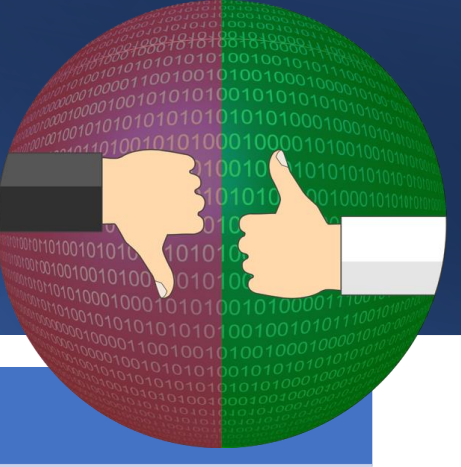
- **Unlabeled Dataset**:
  - ❑This type of dataset comprises data points without any associated explanations, classes, or tags. For instance, images of animals are provided without any information indicating the type of animal in each image. Unlabeled data is typically used in unsupervised learning where the goal is to identify patterns or structures within the data without prior knowledge.
- **Labeled Dataset**:
  - ❑In contrast, a labeled dataset contains data points that are each associated with a specific label, class, or tag that provides information about the data. Using the same example, images of animals would be labeled with tags such as "cat" or "dog" to indicate the content of the image. Labeled data is essential for supervised learning, where an AI model learns to predict the labels of new data based on the patterns it has learned from the labeled training data.



Source

Differentiate Good Data from Bad Data

Source



Good Data:

- Abundant Samples: Good data is characterized by having a large number of data points, which can provide a more comprehensive view of the phenomenon being studied.
- Unbiased: It is representative of the population or phenomena without any inherent biases that could skew results.
- No Missing Data Points: Good data sets are complete or have minimal missing values, which contribute to more accurate and reliable analyses.
- Relevant Features: It contains features (variables) that are important and relevant to the analysis or predictive modeling tasks at hand.
- No Duplicate Samples: Good data is devoid of duplicates, ensuring that each data point is unique and the analysis is not skewed by overrepresentation.

Bad Data:

- Few Samples: Bad data often suffers from a lack of enough data points, which can lead to overfitting in machine learning and unreliable conclusions in analysis.
- Biased: If data is biased, it does not accurately represent the population or phenomenon, leading to biased outcomes and decisions.
- Missing Data Points: Bad data often contains gaps or missing values, which can introduce significant uncertainty or bias into the results.
- Irrelevant Features: It includes features that are not useful for the analysis, which can complicate the modeling process and potentially reduce model performance.
- Duplicate Samples: Bad data may have duplicates, which can distort analytical results and affect the performance of predictive models.

Data Labeling

- **Requirement of Large Labeled Datasets**: Contemporary deep learning models, especially those considered state-of-the-art, rely heavily on extensive volumes of labeled data. The size of the dataset can significantly impact the performance of these models, as they need substantial amounts of data to learn the intricate patterns necessary for tasks such as image and speech recognition, natural language processing, and predictive analytics.
- **Need for a Large Human Labeling Team**: To create these large labeled datasets, a considerable workforce is often necessary. Human labelers are employed to manually annotate and label the data, which can be a time-consuming and labor-intensive process. This is particularly true for tasks that require a high level of precision and cannot be easily automated.
- **Inherent Biases**: Human involvement in data labeling can introduce biases, whether from individual perceptions, cultural contexts, or misunderstanding of the labeling task. Such biases can adversely affect the model's performance, as it may learn and perpetuate these biases. Controls and standardization are necessary to minimize this risk, ensuring that the labels are as objective and accurate as possible.
- **Time Consumption for Data Scientists**: Data curation and labeling can consume a significant portion of data scientists' time. Before even beginning to develop and train models, data scientists often find themselves spending a great deal of time preparing the data—cleaning, labeling, and ensuring it is suitable for use in model training. This highlights the need for efficient data labeling processes and tools that can help streamline this aspect of the data scientist's role.

JSON Lines format

```
{"id": 1, "name": "John Doe"}
{"id": 2, "name": "Jane Doe"}
{"id": 3, "name": "John Smith"}
```

Source

- The JSON Lines format, often abbreviated as JSONL, is a convenient format for storing structured data, allowing it to be processed one record at a time. This can be particularly useful when dealing with large datasets or streaming data, where it's beneficial to read, process, and write data incrementally rather than loading the entire dataset into memory.
- In the context of the JSONL provided:
  - Each line is a complete JSON object.
  - Every line represents a separate data record.
  - This format is line-delimited, which means each new record is placed on a new line.

{“Image 1”: “Cat”}
{“Image 2”: “Dog”}
{“Image 3”: “Lion”}

- JSON Lines is particularly suited for data labeling as it allows each image or data point to be paired with its label in a format that's easy to read and write programmatically.

Manifest Files

Source

```
<?xml version="1.0" encoding="utf-8"?>
<omexManifest xmlns="http://identifiers.org/combine.specifications/omex-manifest">
    <content location="."
        format="http://identifiers.org/combine.specifications/omex"/>
    <content location="models/model.xml"
       format="http://identifiers.org/combine.specifications/sbml"/>
    <content location="simulation.xml" master="true"
       format="http://identifiers.org/combine.specifications/sed-ml"/>
    <content location="doc/article.pdf"
       format="http://purl.org/NET/mediatypes/application/pdf"/>
    <content location="metadata.rdf"
        format="http://identifiers.org/combine.specifications/omex-metadata"/>
</omexManifest>
```

- "Manifest Files," which are used in the context of Amazon SageMaker GroundTruth jobs for labeling data in machine learning projects, particularly in supervised learning tasks such as image classification:
- **Purpose of Manifest Files**: They are used in computer vision classification problems to store data about both the inputs (like images) and their corresponding outputs (labels). This is crucial for supervised machine learning models that require labeled data to learn from.
- **Structure of Manifest Files**: The files are structured in JSON Lines (JSONL) format, meaning each line in the file is a valid JSON object. This format is advantageous for processing large datasets because it allows for reading and writing one record at a time, which can be more efficient and require less memory than processing an entire JSON array.
- **Content of a Manifest File**: Each line in a manifest file corresponds to an individual data record, such as an image, and includes labeling information. This might include the source reference (such as the image's location on S3), metadata about the label (like the class name and confidence level), and additional information like whether the label was human-annotated.
- **Manifest File Data Points**: Each manifest file contains 'N' JSON objects, where 'N' is the total number of images or data points that have been used in the dataset for the GroundTruth job.

Labeling Project


020.jpg


019.jpg


018.jpg


017.jpg


016.jpg


015.jpg


014.jpg


013.jpg


012.jpg


011.jpg


010.jpg


009.jpg


008.jpg


007.jpg


006.jpg


005.jpg

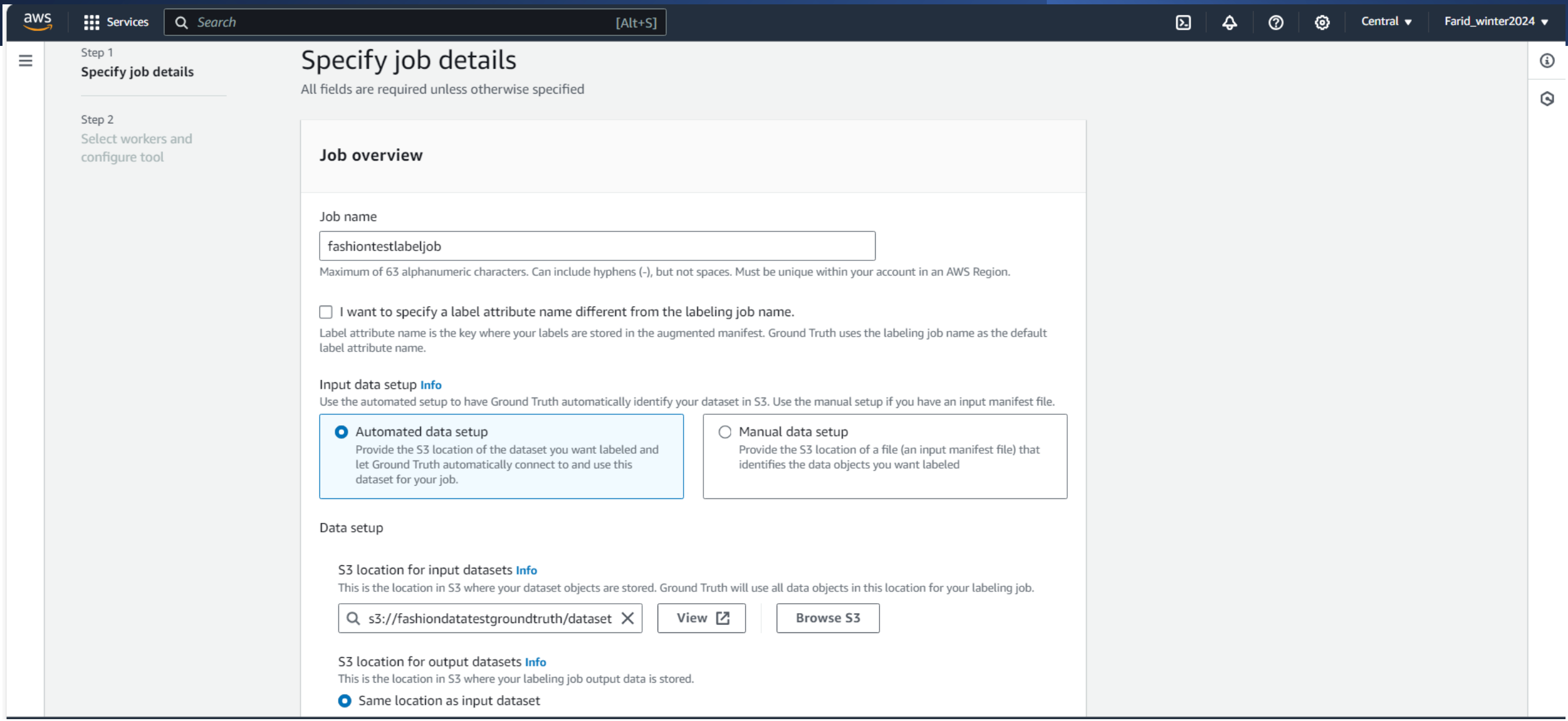

004.jpg


003.jpg

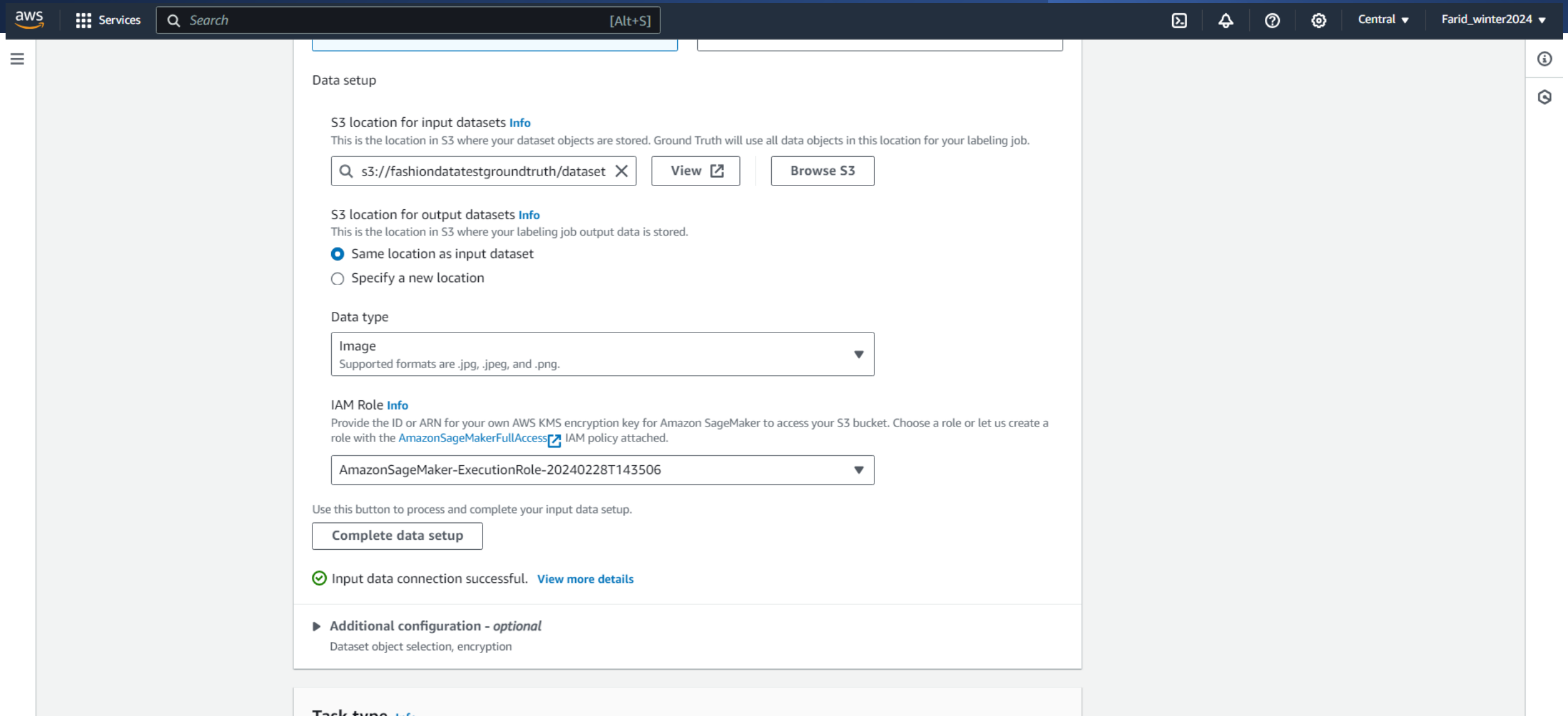

002.jpg


001.jpg

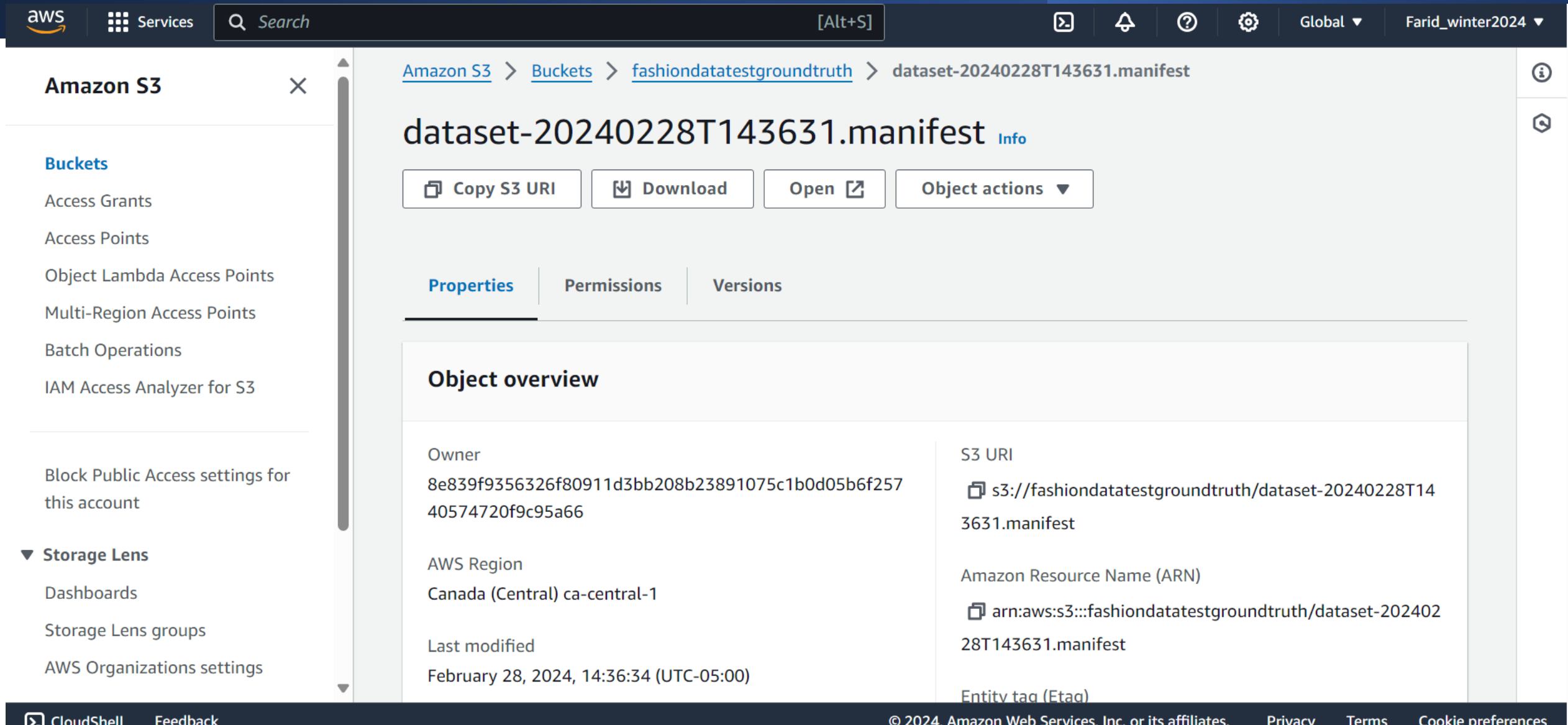
AWS SageMaker GroundTruth



AWS SageMaker GroundTruth



AWS SageMaker GroundTruth



AWS SageMaker GroundTruth


```
{"source-ref":"s3://fashiondatatestgroundtruth/001.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/002.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/003.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/004.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/005.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/006.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/007.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/008.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/009.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/010.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/011.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/012.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/013.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/014.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/015.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/016.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/017.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/018.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/019.jpg"}
{"source-ref":"s3://fashiondatatestgroundtruth/020.jpg"}
```

AWS SageMaker GroundTruth

Select the type of data being labeled to view available task templates for it or select 'Custom' to create your own.

Image

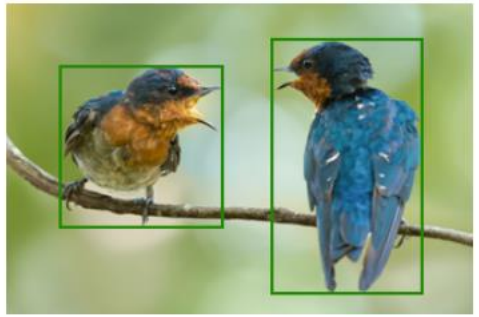
Task selection

Select the task that a human worker will perform to label objects in your dataset.

- Image Classification (Single Label)
  Get workers to categorize images into individual classes. Info

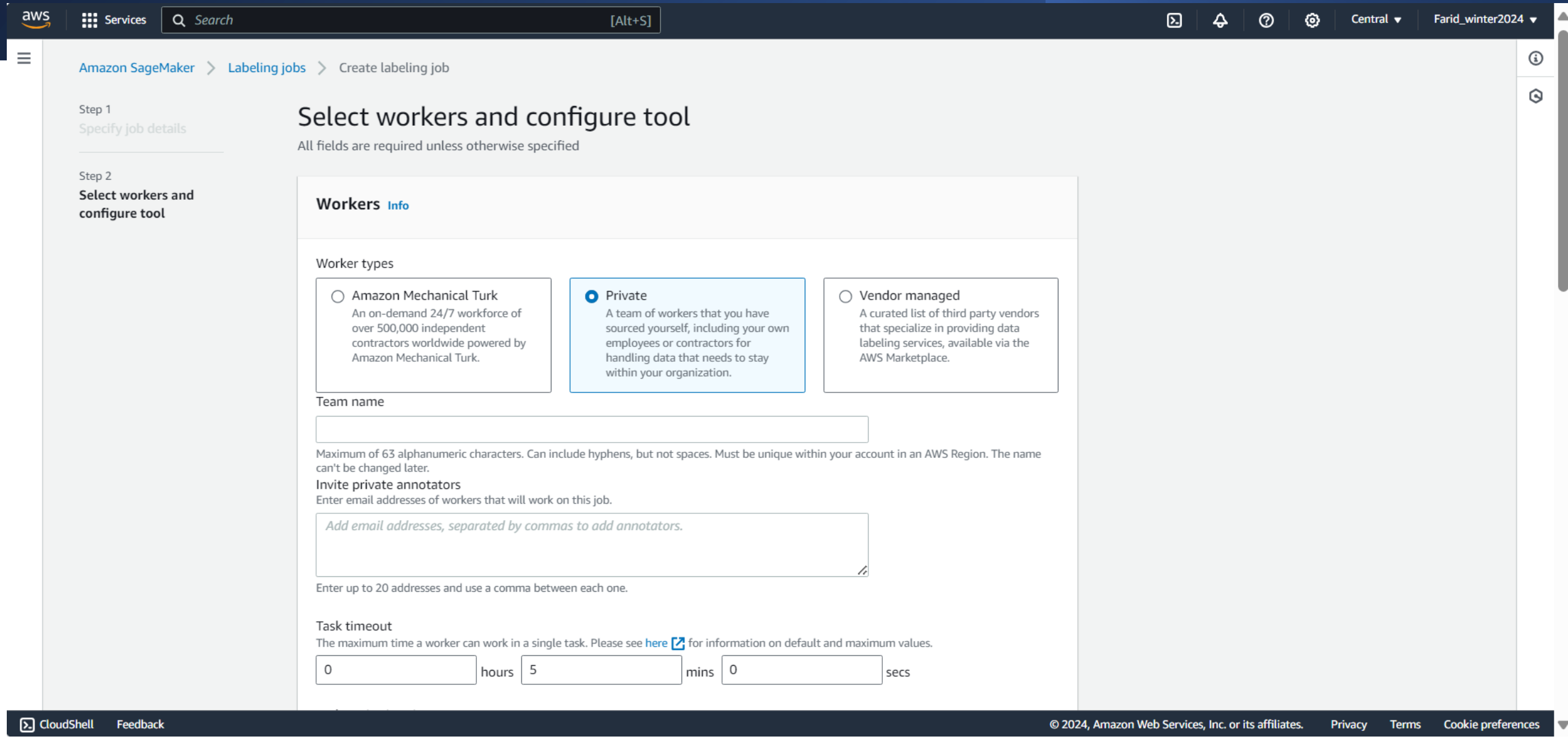
- Image Classification (Multi-label)
  Get workers to categorize images into one or more classes. Info

- Bounding box
  Get workers to draw bounding boxes around specified objects in your images. Info



- Semantic segmentation
  Get workers to draw pixel level labels around specific objects and segments in your images. Info

AWS SageMaker GroundTruth



AWS SageMaker GroundTruth

aws | Services | Search [Alt+S] | Central | Farid_winter2024

Image classification (Single Label) labeling tool

Preview

Provide labeling instructions with examples below for workers. Workers will be viewing these instructions when they perform your task. Workers can choose up to 30 labels. See guidelines for See guidelines for creating high-quality instructions

H1 H2 B I A

Good example

Enter description to explain the correct label to the workers

Add image here

Bad example

Enter description of an incorrect label

Add image here

Enter a brief description of the task

Enter a brief description of the task

Select an option

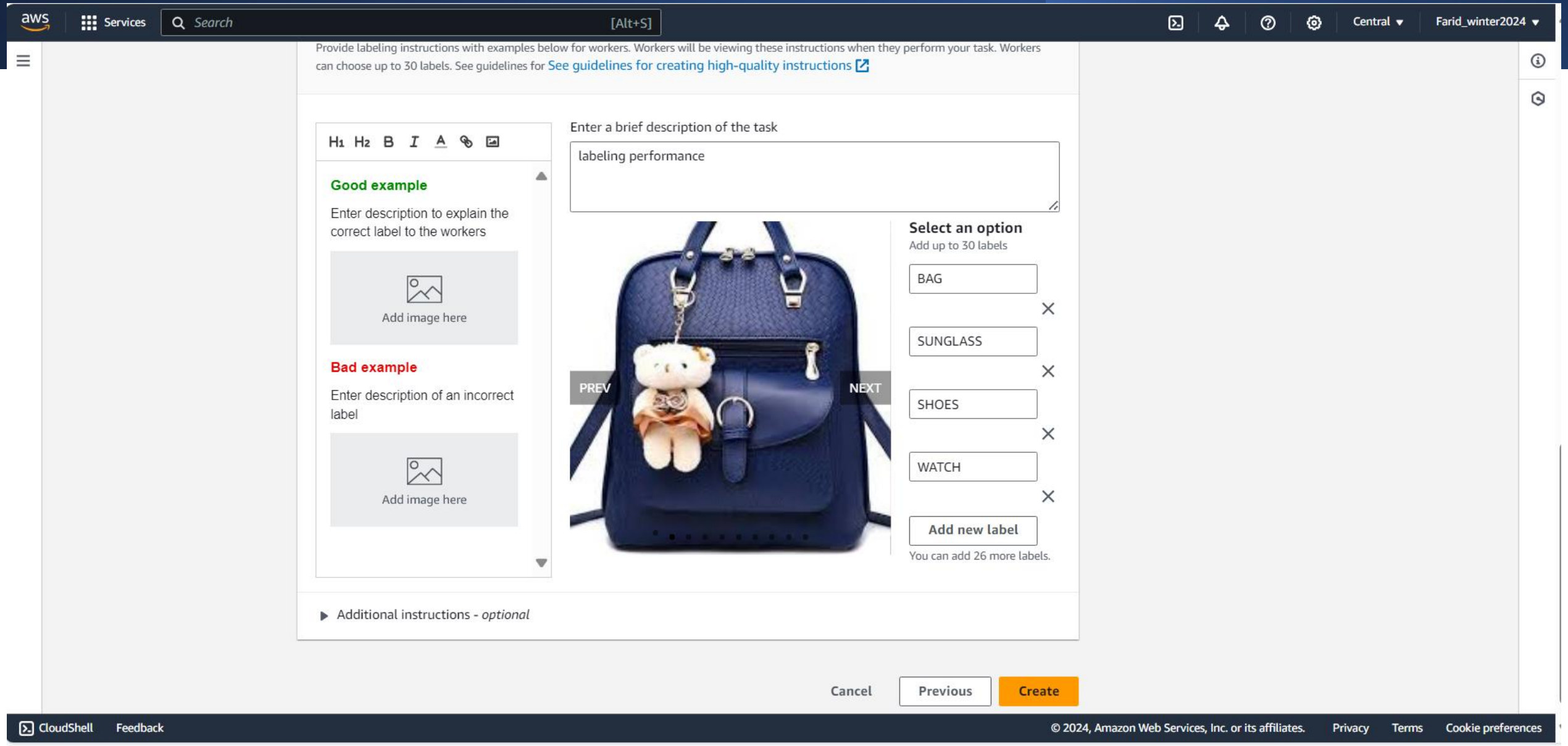
Add up to 30 labels

Add new label

You can add 28 more labels.

▶ Additional instructions - *optional*

AWS SageMaker GroundTruth



AWS SageMaker GroundTruth

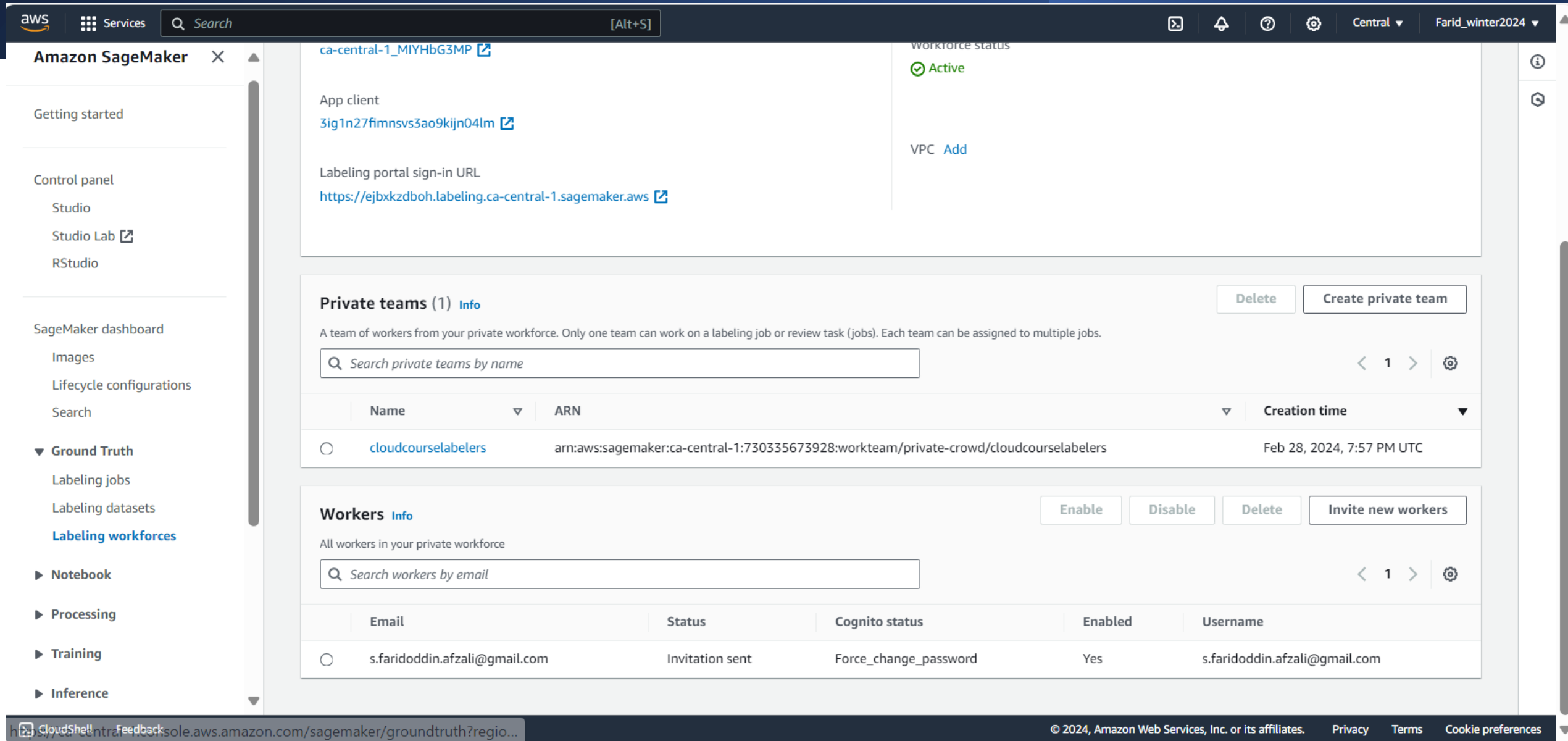
Previewing Answers Submitted by Workers
This message is only visible to you and will not be shown to Workers.
You can test completing the task below and click "Submit" in order to preview the data and format of the submitted results.

Instructions | Shortcuts | labeling performance

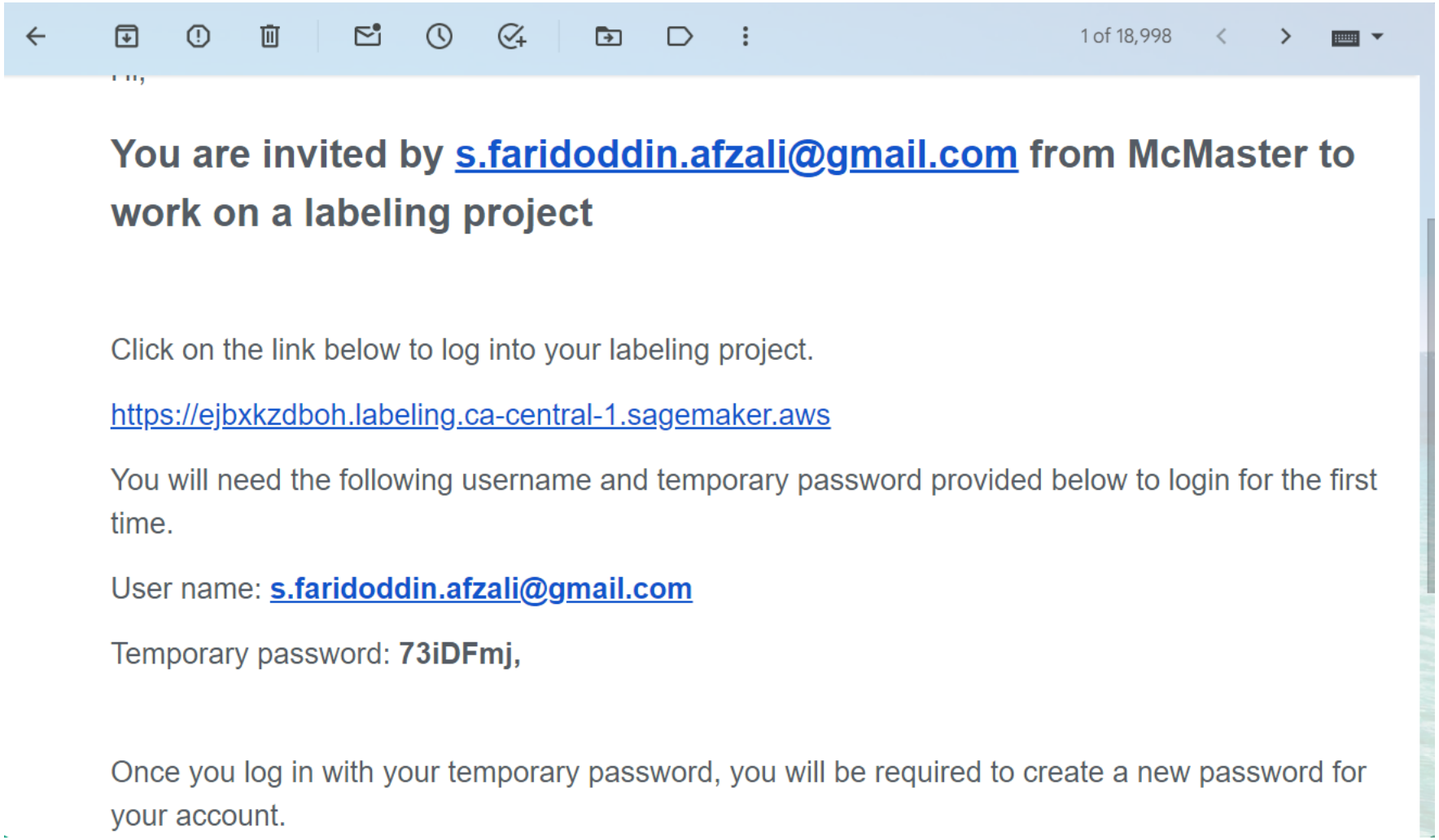
Select an option

BAG	1
SUNGLASS	2
SHOES	3
WATCH	4

AWS SageMaker GroundTruth



AWS SageMaker GroundTruth



AWS SageMaker GroundTruth

Hello, s.faridoddin.afzali@gmail.com

Log out

Show instructions

Jobs (1)

Start working

1

	Task title	Customer ID	Status	Creation time
◉	Image Classification (Single Label): labeling performance	730335673928	Available	February 28, 2024 19:57:15 UTC

AWS SageMaker GroundTruth

Hello, s.faridoddin.afzali@gmail.com | Customer … | Task description: Cate… | Task time: 0:23 of 4 Min 59 Sec | Decline task | Release task | Stop and resume later

Instructions | Shortcuts | labeling performance

Select an option

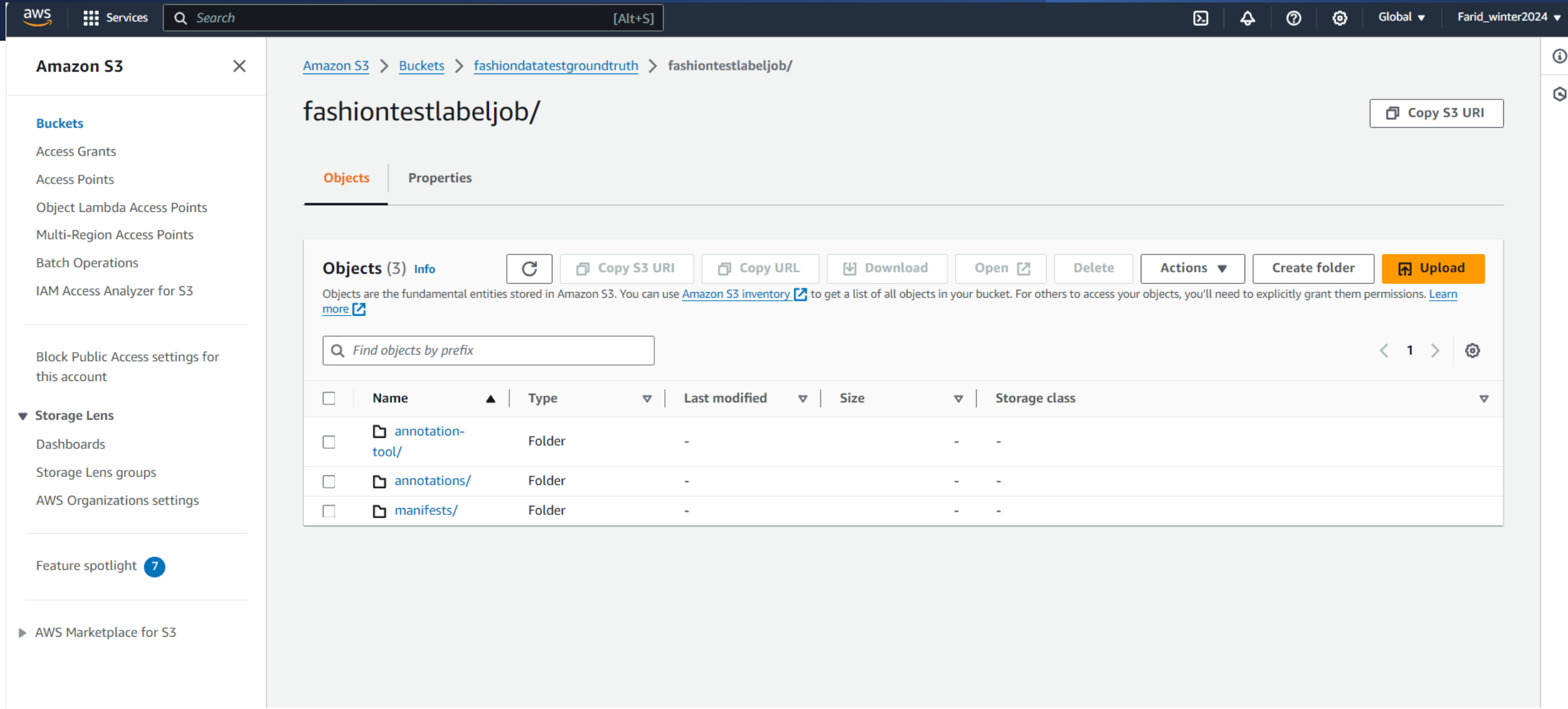
BAG	1
SUNGLASS	2
SHOES	3
WATCH	4

z o m f

Submit

Treat the data in this task as confidential.

AWS SageMaker GroundTruth



AWS SageMaker GroundTruth


```
{"source-ref":"s3://fashiondatatestgroundtruth/001.jpg","fashiontestlabeljob":0,"fashiontestlabeljob-metadata":{"class-name":"BAG","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:12:02.829480"}}
{"source-ref":"s3://fashiondatatestgroundtruth/002.jpg","fashiontestlabeljob":0,"fashiontestlabeljob-metadata":{"class-name":"BAG","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:10:56.312229"}}
{"source-ref":"s3://fashiondatatestgroundtruth/003.jpg","fashiontestlabeljob":0,"fashiontestlabeljob-metadata":{"class-name":"BAG","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:12:02.829471"}}
{"source-ref":"s3://fashiondatatestgroundtruth/004.jpg","fashiontestlabeljob":0,"fashiontestlabeljob-metadata":{"class-name":"BAG","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:12:02.829516"}}
{"source-ref":"s3://fashiondatatestgroundtruth/005.jpg","fashiontestlabeljob":0,"fashiontestlabeljob-metadata":{"class-name":"BAG","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:10:56.312237"}}
{"source-ref":"s3://fashiondatatestgroundtruth/006.jpg","fashiontestlabeljob":0,"fashiontestlabeljob-metadata":{"class-name":"BAG","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:12:02.829454"}}
{"source-ref":"s3://fashiondatatestgroundtruth/007.jpg","fashiontestlabeljob":1,"fashiontestlabeljob-metadata":{"class-name":"SUNGLASS","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:10:56.312209"}}
{"source-ref":"s3://fashiondatatestgroundtruth/008.jpg","fashiontestlabeljob":1,"fashiontestlabeljob-metadata":{"class-name":"SUNGLASS","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:10:56.312296"}}
{"source-ref":"s3://fashiondatatestgroundtruth/009.jpg","fashiontestlabeljob":1,"fashiontestlabeljob-metadata":{"class-name":"SUNGLASS","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:12:02.829463"}}
{"source-ref":"s3://fashiondatatestgroundtruth/010.jpg","fashiontestlabeljob":1,"fashiontestlabeljob-metadata":{"class-name":"SUNGLASS","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:10:56.312183"}}
{"source-ref":"s3://fashiondatatestgroundtruth/011.jpg","fashiontestlabeljob":2,"fashiontestlabeljob-metadata":{"class-name":"SHOES","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:12:02.829444"}}
{"source-ref":"s3://fashiondatatestgroundtruth/012.jpg","fashiontestlabeljob":2,"fashiontestlabeljob-metadata":{"class-name":"SHOES","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:12:02.829489"}}
{"source-ref":"s3://fashiondatatestgroundtruth/013.jpg","fashiontestlabeljob":2,"fashiontestlabeljob-metadata":{"class-name":"SHOES","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:12:02.829420"}}
{"source-ref":"s3://fashiondatatestgroundtruth/014.jpg","fashiontestlabeljob":2,"fashiontestlabeljob-metadata":{"class-name":"SHOES","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:10:56.312260"}}
{"source-ref":"s3://fashiondatatestgroundtruth/015.jpg","fashiontestlabeljob":2,"fashiontestlabeljob-metadata":{"class-name":"SHOES","job-name":"labeling-job/fashiontestlabeljob","confidence":0,"type":"groundtruth/image-classification","human-annotated":"yes","creation-date":"2024-02-28T20:12:02.829498"}}
```